THIS WEEK

EDITORIALS

ECONOMICS Is it possible for ecology to offer a natural wealth service? **p.266**

WORLD VIEW The world won't pay, so the forest must **p.267**

PANDAS Old trees could be key to conservation **p.268**

Smallpox should be saved

Secure virus stocks in the United States and Russia may still prove useful and should not be destroyed. A political compromise is the best way to make that happen.

For much of the world's population, smallpox is a disease of history. The variola virus that causes it last swept through humans in a natural outbreak in Somalia in 1977, and the world was declared free of smallpox in 1980. The disease that killed Queen Mary II of England, Tsar Peter II of Russia, King Louis XV of France and hundreds of millions more during the past century alone is gone — but not forgotten. Smallpox lives on in the memories of those who witnessed its awful impact first hand. It is a terrifying spectre for those who warn that terrorists may seek to spread disease. And it survives in two laboratories, where research continues on live virus.

The fate of these known virus stocks, held in secure laboratories in the United States and Russia, is once again the subject of debate. In May 2011, the World Health Assembly — the decision-making body of the World Health Organization — will vote on whether to set a date by which these collections should be destroyed. They should not be eliminated, at least not completely.

This journal has previously argued that, because the possibility cannot be ruled out that other, secret stocks of smallpox are being held elsewhere, the benefits of continued access to live virus stocks for research outweigh the risks of maintaining them. In 1999 (*Nature* **398,** 733; 1999), we wrote: "Rightly, public-health advocates bemoan the prospect of any measure that increases the risk of a re-emergence of this scourge. But, given the impossibility of knowing who now possesses the virus, and from where it might appear, it is better to have a number of arrows in the quiver than to destroy the stock and cross our collective fingers." The world has changed much since then, and the US terrorist attacks of September 2001 and the subsequent focus on prospects for bioterrorism have increased the stakes further. Smallpox would be an effective weapon — it spreads easily and kills almost one-third of the people it infects. Furthermore, the triumph of smallpox eradication after widespread vaccination in the 1960s and 1970s means that some 40% of the world's population has no immunity.

In an essay published online earlier this month in the journal *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, security expert Jonathan Tucker argues that many of the stated goals of the World Health Organization's smallpox research programme have been achieved — such as antiviral drugs and diagnostic tools — and that there is a diminished need for live virus to be retained (J. B. Tucker *Biosecur. Bioterror*. doi:10.1089/bsp.2010.0065; 2011). This may be true, but further study of the virus could still reveal a huge amount, both on the specifics of what makes it such a formidable foe and on human immunology and viral pathogenesis in general.

The scientific case for retaining live variola virus to improve public health is strong. The risk of doing so is largely political. The threat of an accidental release or theft from the biosecure repositories at which it is held — at the Centers for Disease Control and Prevention in Atlanta, Georgia, and the State Research Center of Virology and Biotechnology ('Vector') near Novosibirsk — seems remote. Destruction of the stocks would be a largely symbolic step.

Countries in Africa and Asia endured the havoc of smallpox most recently, and fear of a repeat outbreak — as well as mistrust of those who control the stocks — has produced strong calls from these nations for its destruction. There is equally strong insistence in countries such as the United States and Russia that it should be retained for defensive research. In a recent article in *The New York Times*, former US government heavy-hitters Kenneth Bernard and Richard Danzig call elimination of the virus stocks a "bad idea". They say: "We have an obligation to ensure that our children's children never have to worry about seeing this ancient scourge kill yet again. Destroying the remaining research specimens of smallpox at this time will prevent us from fulfilling this obligation."

"Smallpox is a disease of history, but it cannot be consigned to the past."

Tucker warns that the result of this political stand-off, further complicated by the rare involvement of security and defence officials in a public-health matter, could be a "diplomatic train wreck" — particularly given that the World Health Assembly will seek a consensus decision. As part of a compromise solution, he suggests that the United States and Russia should agree to destroy part of their collections, such as hybrid viruses that are scientifically redundant. A handful of strains would continue to be held at each site, for research in the short term and reference afterwards. That seems a sensible approach. True, the world will be denied a final victory in the heroic effort to defeat smallpox, but uncertainty over undeclared surviving stocks would make that a hollow accomplishment anyway. So, too, would development of molecular-biology techniques that could see the smallpox virus reconstructed. Smallpox is a disease of history, but it cannot be consigned to the past. ■

Goodbye Tevatron

The decision to shut down the ageing particle collider at Fermilab is the right one.

On 10 January, the US Department of Energy (DOE) announced that the United States will close its only particle collider, the Tevatron, located at Fermilab in Batavia, Illinois. The decision to shutter the decades-old collider this year (see *Nature* **469,** 141; 2011) goes against the recommendations of the US particle-physics community's scientific advisory group, which believes that the machine still has work to do. But given the financial constraints facing Fermilab, and its ambitions for the future, the choice is at once courageous and correct. Moreover, it should serve as an example to other scientific communities that are clinging to ageing

infrastructure at the expense of the next generation.

As is often the case in high-energy physics, the success of the Tevatron can be measured in orders of magnitude. Since the Tevatron's opening in 1985, its luminosity, a measure of the number of collisions produced, has increased by a factor of 20 million. Its detectors have gathered data at an exponential rate, along the way discovering the top quark in 1995 and the tau neutrino in 2000.

But few discoveries have emerged from the accelerator in the past decade. The search for heavier particles, such as the Higgs boson, a critical part of the mechanism for mass, uses higher luminosities and energies than the Tevatron could ever produce. To make up for its lack of power, the machine has been forced into a war of attrition: running for years beyond its expected lifetime in the hope of collecting enough data to spot something new. It is an inefficient way to do science.

Despite its mounting obsolescence, scientists have continued to push to extend the Tevatron's run. In the beginning, the case was made that it could yet beat its European rival, the Large Hadron Collider (LHC) at CERN, Europe's particle-physics lab near Geneva, Switzerland, to the discovery of the Higgs. That argument held some weight while the European collider was facing early teething pains, but last year the LHC roared to life and is now churning out data at a staggering rate (see page 282). The latest scientific case for keeping the Tevatron running until the end of 2014 revolves around its ability only to plug gaps in the LHC's work.

Although scientifically persuasive, this argument must be measured against the US$35 million needed each year to operate the machine. That sum might seem small in comparison to the US research budget as a whole, but it represents roughly 10% of Fermilab's annual operating funds. Lab leaders would like to put the money elsewhere: faced with the truth that the Tevatron cannot run forever, they are hoping to develop a strong programme in neutrino physics. The new direction will keep the scientists at Fermilab working for years to come and could even lead to future collider projects.

To administrators at Fermilab and the DOE, the decision was clear: the lab must shut down the Tevatron in order to move forward. In one sense it is a risky strategy — to trade a steady income for the promise of better to come. Funds for the new programme are far from guaranteed and could yet be a tall order in the current economic climate. But in making the sacrifice, the lab showed that it is serious about its new neutrino programme. It is also providing its political supporters with evidence that it is acting as a responsible research body, something that may help them win the additional money needed for the new programme.

"It can serve as an example to scientific communities that are clinging to ageing infrastructure at the expense of the next generation."

The move stands in stark contrast to other fields that have chosen to protect the old over the new. Some in the astronomy community have lobbied hard to keep 'historic' facilities such as the Arecibo Observatory in the United States and Jodrell Bank in the United Kingdom. Elsewhere in Europe, geriatric research reactors and synchrotron light sources are kept running more out of national pride than scientific need.

The case for closing many of these facilities is less clear-cut than the Tevatron, however. The collider has obviously been superseded by its European rival, and high-energy physicists have little choice but to move on. By contrast, telescopes and synchrotrons can always gather more data on another star or a new material. Yet at some stage, the scientific value of these facilities becomes so diminished that it must be weighed against the new opportunities the funding could give to the community. Rather than fighting protracted political battles against their inevitable closure, researchers should provide new ideas and facilities that could be leveraged using funds from the closure of the last generation. Such a strategy is more risky, but it will ultimately prove much more rewarding. ■

Natural wealth

Ecological models can be used to guide economic policy — but should they?

The first law of economists states that for every economist there exists an equal and opposite economist. In that spirit, *Nature* this week presents related articles that take contradictory stances on what the banking system could, or indeed should, learn from ecologists. In a Perspective (see page 351), Andrew Haldane, executive director of financial stability at the Bank of England in London, and Robert May, a theoretical ecologist at the University of Oxford, UK, and former chief scientific adviser to the UK government, argue that the stability of complex networks of linked dependencies that make up the world's financial system can be tested using a simplified ecological model. They say that the results of their approach highlight ways in which economic policy could be changed to make the system more secure. Regulation of elements such as liquidity ratios, banking structures and trade in complex derivatives, they say, should work to protect the system rather than individual banks.

In a News & Views Forum (see page 302), this approach is criticized as potentially dangerous by Neil Johnson, a physicist at the University of Miami in Florida who studies real-world complex systems. Johnson says that policy implications drawn from such a simplistic model will be unreliable, because they will depend so heavily on the assumptions used to prepare the model. But Thomas Lux, an economist at the University of Kiel in Germany, offers a more supportive view. He says that the similarities between financial markets and ecology are real and relevant. For example, the 2008 default of the Lehman Brothers financial services firm had contagious effects similar to those of a 'super-spreader' of disease. Researchers should go beyond Haldane and May's simple model, says Lux, and build in factors such as interbank credit lines.

Given such contrasting views, how seriously should policy-makers take lessons from nature when it comes to financial regulation? Using natural models to set policy has proven difficult, even for ecological policy. Witness the controversy that has surrounded the UK government's handling of foot-and-mouth outbreaks and whether to tackle bovine tuberculosis by culling badgers, which can spread the disease. In other areas, such as conservation and fisheries, research offers more clear-cut recommendations. But in such cases there is often a failure of political will to follow through. Which is the case in finance policy? Perhaps the best approach is to keep a cautiously open mind — or to remember that the second law of economists states that both the economists cited in the first law are wrong.

Wider discussion of the ideas raised by Haldane and May will inevitably throw up many questions. Among them, we expect, will be: why is *Nature* publishing research in economics? As a natural-sciences journal, most of economics falls outside our primary-research remit, and there are some fundamental differences in approach between the fields. For example, in financial systems, as Haldane and May point out, evolutionary forces often act for the survival of the fattest, rather than the fittest. But in this case, we think the clear links that the authors draw with ecology will make their study of interest to readers. More generally, we are happy to publish primary economic research that has significant scientific implications, for example in fields such as behaviour, conservation biology, systems biology or physics. With such an approach, perhaps both economic science and natural science can benefit. ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv

L. SEMPERTEGUI

Day of reckoning for Ecuador's biodiversity

The world's indifference to a request for $3.6 billion to preserve a diversity hot spot may push the country to extract oil there, says **Kelly Swing**.

How much is the world willing to pay to protect a piece of Amazonian Ecuador that is home to the greatest concentration of species on the planet? Ecuador's government has asked for US$3.6 billion to refrain from developing the region, in the Yasuní Biosphere Reserve, but the reaction of the world suggests that the price is too high. We need to accept that big oil firms are coming to Yasuní, and the struggle now must be how to limit the damage, rather than prevent it.

The effort to save the region mirrors the plot of 2009's blockbuster film *Avatar*. Yasuní is the set and its ancestral people are the Waorani. The precious commodity beneath their feet is fossil fuel. The region and its people are highly threatened by oil extraction and the collateral damage that such development brings. For most of the past half-century, Ecuador's economy has depended heavily on oil revenue. Oil-field development continues to push into more remote areas to keep the cash flowing. It has already reached into the heart of the Yasuní reserve.

In an innovative proposal set out in 2007, President Rafael Correa offered to sell the world a guarantee that oil would not be extracted from the Ishpingo-Tambococha-Tiputini (ITT) block, a 1,200-square-kilometre portion of Yasuní on the country's eastern border with Peru. The $3.6-billion price is about half of what Ecuador could expect to make from the roughly 850 million barrels of crude oil reckoned to be there. In exchange, Correa said that the area would be legally set aside, and become off-limits to extraction in perpetuity. In Ecuador, the plan is optimistically referred to as Plan A.

Two decades in Yasuní have infused me with an intense desire for this initiative to work. Unfortunately, after three years of worldwide campaigning, it is heartbreaking to say that I don't see it happening. All the environmentalists who see the initiative as Yasuní's last chance don't want to believe it might fail. Out of fear that we might jinx it, most of us haven't even been willing to hint at its possible failure. But the trust fund set up to receive the money contains just $100,000 so far. Correa says he will evaluate its accumulation rate in a few months. If more money does not come, he could withdraw the offer.

Why the poor response? Some people suggest that his initiative is essentially an ugly ultimatum — either the world pays to save Yasuní or Ecuador will have no choice but to exploit the oil beneath it. In this scenario, the Ecuadorian government can blame the rest of the world for any negative consequences. Ecuador's vice-president, Lenín Moreno, denies the reserve is being held for ransom, and points out that his country would be the greatest contributor, as it is willing to accept losses of $3.6 billion to make the plan work. Either way, there is a credibility gap. According to stipulations, Ecuador's responsibility would be to use contributions for sustainable development that would decrease the country's energy footprint and boost conservation efforts. Does the world believe that Ecuador can be trusted to use the funds in this way and to keep its promises — not only for the next decade but forever? Realistically, no conservation plan can demand that any sovereign nation be permanently deprived of rights to develop and profit from its own natural resources.

IF ROADS CONTINUE TO BE BUILT TO PROVIDE ACCESS THEN, WITHOUT QUESTION, ALL IS LOST.

We have to realize that Plan B, to extract and sell the oil, is ready to be set in motion. Historically, such development north of the Napo River has led to widespread oil spills and surface dumping of formation water, and the opening of access roads has been followed by waves of settlers, deforestation, expansion of the agricultural frontier, acculturation of indigenous groups, timber harvest and the marketing of bushmeat, resulting in near complete destruction of that region. This same process has already compromised the accessible western parts of Yasuní.

If the crude oil in the ITT block and surrounding region is exploited using the same methods, the result will be nothing short of a world-class environmental disaster. If an earthquake, hurricane or bomb caused the same amount of destruction in an instant, no one could ignore the impact, or would simply stand by and accept it. Because this tremendous loss would be stretched out over decades, few are likely to notice the gradual change.

Does this have to be the future? Oil companies say they can do better, if given the chance. The oil industry has improved its strategies and technologies to minimize risks, but minimizing risks is not the same as eliminating them. Just think back on BP's summer of oil in the Gulf of Mexico.

If extraction is deemed necessary and unavoidable in the ITT (and probably most of the Yasuní reserve), any plan must require that the job genuinely be done in an environmentally sound way. This can be done only if the oil is extracted without new roads, which open the way to so many destructive forces. If roads continue to be built to provide access then, without question, all is lost. Instead, the oil must be extracted using 'off-shore' strategies, which are more expensive but do less damage. It goes without saying that any operations would have to include proper maintenance and independent monitoring, but having isolated oil platforms scattered across the landscape has to be better than horizon-to-horizon deforestation.

Correa's government has already indicated that it would sacrifice revenues of $3.6 billion to save a small portion of this regional biodiversity hot spot. A simple policy of no more roads built in the reserve would go a long way to protecting much more of Yasuní, well beyond the bounds of the ITT block, and for far less money. ■

➲ NATURE.COM
Discuss this article online at:
go.nature.com/33iqpt

Kelly Swing *is founding director of the Tiputini Biodiversity Station, University of San Francisco de Quito, Ecuador.*
e-mail: kswing@usfq.edu.ec

RESEARCH HIGHLIGHTS

Selections from the scientific literature

STEM CELLS

Robo protein guide for cell transplants

A protein used by blood stem cells to adhere to bone marrow could offer a way to improve the success of blood stem-cell transplants.

Certain illnesses, such as some blood cancers, have long been treated with blood — or haematopoietic — stem-cell transplants, but the procedure is risky. To learn more about how to manipulate these cells, Camilla Forsberg at the University of California, Santa Cruz, and her colleagues studied the Robo4 protein.

They found that, in mice, haematopoietic stem cells lacking Robo4 were fewer in number, and less able to anchor themselves in the bone marrow after transplantation. The authors also revealed that Robo4 works with another protein, Cxcr4, to localize transplanted stem cells to the bone marrow.

Cell Stem Cell 8, **72–83 (2011)**

DLILLC/CORBIS

CONSERVATION BIOLOGY

Pandas prefer their trees mature

Giant pandas are famous for their love of bamboo, but it seems that they also have a preference for old forests, suggesting that maintaining such habitats is important for conserving this iconic endangered species.

Fuwen Wei at the Chinese Academy of Science in Beijing and his colleagues analysed data collected between 1999 and 2003 on 4,908 plots of land in Sichuan province for signs of pandas (*Ailuropoda melanoleuca,* **pictured**). They found that the presence of old-growing forests was just as good an indicator that pandas live in the area as the presence of bamboo.

The authors suggest that maps and models of suitable habitats for pandas should be revised to prioritize old forests. They also call on the Chinese government to renew its ban on logging.

Biol. Lett. **doi:1011098/rsbl.2010.1081 (2010)**

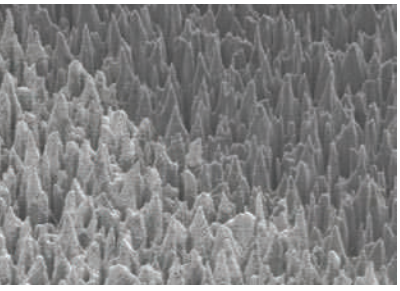
MATERIALS SCIENCE

Gas keeps drag low

The flow of water on solid surfaces is significantly impeded by frictional forces — which is bad news for, say, marine vehicles. A gas layer can be introduced at the solid–liquid interface as a lubricant, but even slight hydraulic pressure can destroy this layer. Choongyeop Lee and Chang-Jin Kim at the University of California, Los Angeles, have devised a way to keep the gas layer intact and cut drag even in underwater conditions.

The duo began with a highly hydrophobic surface studded with 50-micrometre-high pillars and gold-coated nanostructures (**pictured**), and submerged this in water. The gold coating allowed an electrolytic reaction to occur, generating gas at its surface when water made contact. Bubbles formed only in areas where there had been no gas before, and because of the surface's architecture, the bubbles spread uniformly across the surface.

Phys. Rev. Lett. 106, **014502 (2011)**



AM. PHYS. SOC.

IMMUNOLOGY

Stronger shields against the flu

People's immune responses to infection with the 2009 H1N1 influenza pandemic virus included neutralizing antibodies effective against many flu virus strains.

Rafi Ahmed at Emory University in Atlanta, Georgia, Patrick Wilson at the University of Chicago in Illinois and their group analysed 86 antibodies made by immune cells from nine patients infected with H1N1. Surprisingly, 63% of the antibodies bound to flu viruses that had appeared before the 2009 pandemic. Some even worked against the 1918 pandemic virus and the H5N1 avian influenza virus. Three antibodies were generated from patients' cells and tested in mice. All three protected the animals against a lethal dose of H1N1, whether given before or after exposure. Such broadly protective antibodies — long sought, but elusive — could aid the design of a 'universal' flu vaccine, the team speculates.

J. Exp. Med. **doi:10.1084/jem.20101352 (2011)**

ASTRONOMY

The dance of three stars

Researchers have spotted a cosmic ballet — two stars spinning around one another while both orbiting a larger star. The trio, named KOI-126,

provided a rare opportunity to precisely measure the masses of binary stars.

The triple system is about 1,000 parsecs from Earth. The smaller duo have masses of about one-fifth and one-quarter that of the Sun, and orbit one another in 1.76 days. They take about 34 days to go around their larger partner, which has a mass 1.3 times that of the Sun. Joshua Carter of the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts, and his team calculated the masses using the space-based Kepler telescope by observing the dimming of light as the binary stars passed in front of the larger star. They produced a model that fitted the data to determine the dynamics and characteristics of the low-mass pair.

Science **doi:10.1126/science.1201274 (2011)**

NEUROSCIENCE

Ringing in the brain

People with tinnitus hear a ringing even when there is no sound. The prevailing theory blames a malfunctioning auditory system. However, it seems that abnormalities in the brain's limbic regions, which determine which sensations are important and how they are experienced, may also be involved.

Josef Rauschecker and his colleagues at Georgetown University in Washington DC used functional magnetic resonance imaging to scan the brains of 22 volunteers, half of whom had tinnitus, while they listened to various sounds. Patients with tinnitus showed heightened activity in the nucleus accumbens — a key limbic region — when presented with sounds that matched the frequency of the 'ringing' in their ears. They also had anatomical differences in the ventromedial prefrontal cortex, another limbic area.

The researchers suggest that an abnormal limbic system elevates the perceived importance of the tinnitus sound or fails to suppress it, and that interactions between the auditory and limbic systems may be at the root of this disorder.

Neuron 69, **33–43 (2011)**

DRUG DEVELOPMENT

Promoter predicts drug results

A clinical trial of a drug to treat behavioural features of the genetic disorder fragile X syndrome has revealed a genetic signature that seems to predict which patients will respond to the treatment.

Fragile X syndrome results from excessive methylation of a regulatory section — called the promoter — of the *FMR1* gene. This silences the gene and increases signalling in the pathway for the brain receptor mGluR5. The drug candidate, AFQ056, blocks mGluR5 activation.

Baltazar Gomez-Mancilla at the Novartis Institutes for Biomedical Research in Basel, Switzerland, and his group assessed the drug's effects on the behaviour of 30 men who had either a fully or partly methylated *FMR1* promoter region. Seven with a fully methylated promoter showed behavioural improvements after 19–20 days, whereas the behaviour of those with only partial methylation did not differ significantly between drug and placebo treatments.

Sci. Transl. Med. 3, **64ra1 (2011)**

THEORETICAL PHYSICS

Relativity starts your car

Relativity is generally invoked in lofty thought experiments often involving fast-moving spacecraft, but new work shows that it also applies to the everyday automobile.

Cars are started using lead–acid batteries, which generate energy using electrochemical reactions between lead compounds and sulphuric acid. Rajeev Ahuja of Uppsala University in Sweden and his colleagues modelled the reactions and found that as electrons move at high speed around a lead nucleus, their energy levels change owing to relativity. The authors conclude that the change accounts for 1.7–1.8 volts of a standard 2.13-volt lead–acid cell.

Phys. Rev. Lett. 106, **018301 (2011)**

COMMUNITY CHOICE

The most viewed papers in science

PHYLOGENETICS

How one elephant became two

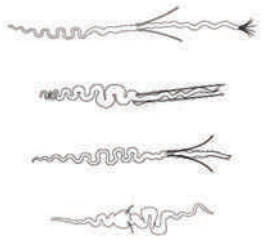
HIGHLY READ on plosbiology.org the weeks beginning 27 Dec and 3 Jan

Whether savannah- and forest-dwelling African elephants belong to the same or different species has long been the subject of debate. An analysis of the creatures' genetic ancestry confirms that the two species are separate.

Nadin Rohland at Harvard Medical School in Boston, Massachusetts, and her co-workers sequenced the nuclear genomes of the forest elephant (*Loxodonta cyclotis*) and the savannah one (*L. africana*), as well as that of the Asian elephant (*Elephas maximus*). They also extracted and sequenced DNA from the extinct woolly mammoth (*Mammuthus primigenius*) and American mastodon (*Mammut americanum*) — ancient ancestors of today's elephants. By comparing all of these genomes, the team found that the forest and savannah elephants diverged into separate species between 7.1 million and 1.9 million years ago — much earlier than previously proposed.

PLoS Biol. 8, **e1000564 (2010)**

For more on this research, see go.nature.com/91a4zy



EVOLUTIONARY BIOLOGY

The many styles of sperm

Of all the cell types in the animal kingdom, the most diverse are sperm, which can be adorned with tails, hairs, bristles and more. It seems that sex drives the evolution of sperm morphology.

Lukas Schärer at the University of Basel in Switzerland and his colleagues watched 16 species of promiscuous hermaphroditic flatworm (*Macrostomum*), which have a variety of sperm shapes (**pictured**), mating under a microscope. After sex, some species suck out the ejaculate, possibly as a way of selecting which sperm are ultimately accepted.

The researchers found that those species that exhibit this sucking behaviour have ornate sperm with features such as a pair of long bristles emerging at the mid-point and a tail resembling a paint brush. These appendages can become lodged in the female orifice after copulation, preventing the sperm from being sucked out. Species that don't remove sperm have evolved simpler sperm that tend to be smaller and lack hairs or bristles.

Proc. Natl Acad. Sci. USA **doi:10.1073/pnas.1013892108 (2011)**

For a longer story on this research, see go.nature.com/8jvjws

NATURE.COM

For the latest research published by *Nature* visit:
www.nature.com/latestresearch

NATL ACAD. SCI.

SEVEN DAYS *The news in brief*

POLICY

Haiti's cholera fight

Health officials have outlined plans for a proposed cholera vaccination programme in Haiti — but there is still tension over how large the effort should be. As the country struggles to recover from the earthquake that devastated it last January, some experts have pushed for a small pilot project, but the government wants a larger programme to be rolled out as soon as possible. See page 273 for more.

NIH conflict clash

The US National Institutes of Health (NIH) should issue regulations governing conflicts of interest for the institutions at which its grantees work, a report by the inspector general of the Department of Health and Human Services urges. The NIH is revamping rules that address the reporting of potential conflicts by individual investigators but it doesn't have analogous rules for universities and medical centres, even though these are required by law, the 10 January report notes. The agency says it is "carefully considering" comments it has received on institutional conflicts as it finalizes changes to the rules for individuals.

French drug reform

France's health minister Xavier Bertrand last week pledged to reform the country's drug-regulation system in the wake of a damning official report into why the weight-loss drug Mediator remained on the market for years despite concerns over potentially lethal side effects. The AFSSAPS, the French state body that approves drugs for marketing, banned the drug only in 2009, even though questions were raised about its impact on heart disease more than a decade ago, a delay that the report says may have contributed to some 500 premature deaths. The AFSSAPS will now investigate 76 other products, Bertrand says.

J. GENTNER/AP

Permit for mountaintop mine revoked

The US Environmental Protection Agency (EPA) has for the first time revoked a permit for mountaintop mining. Arch Coal, a mining company based in St Louis, Missouri, obtained the permit for what would have been the country's largest mountaintop mine in 2007 from the Army Corps of Engineers, which is responsible for developing and maintaining US water resources (see *Nature* **467,** 1021; 2010). But the EPA has ruled that Spruce 1 Mine in the Appalachian Mountains of West Virginia presented major environmental and water-quality concerns, and would jeopardize the health of local communities. US mountaintop mines, such as the one on Kayford Mountain (pictured), are largely found in the Appalachians.

NASA waste

NASA could end up spending US$575 million on a space programme that has already been cancelled, the agency's inspector general has warned Congress. The US government is currently funded by a 'continuing resolution', which requires agencies to fund existing programmes at last year's level until Congress passes a new budget. This means that NASA has to fund Constellation, former president George W. Bush's programme to return to the Moon and reach Mars, at $200 million a month until 4 March. Congress passed legislation to cancel the programme last October. See go.nature.com/pshdef for more.

RESEARCH

Hunger still rife

Entire segments of the food production system have been damagingly neglected in international attempts to reduce hunger and poverty, according to a report from the Worldwatch Institute, an independent research organization based in Washington DC. The 'State of the World' report, published last week, says that previous approaches to feeding the world's population have "not really worked", given that 925 million people worldwide still go hungry every day. Worldwatch calls for new approaches, such as growing a wider variety of crops and reducing reliance on chemical fertilizers. See go.nature.com/bbkrbb for more.

Climate records

Last year was one of the two warmest years on record, according to the US National Oceanic and Atmospheric Administration (NOAA). Only

NUMBER CRUNCH

18%

The number of health advocacy groups receiving funding from Eli Lilly in the first half of 2007 that acknowledged it in their annual reports.

Source: *Am. J. Public Health*

2005 had an equivalent average global surface temperature. Global average temperatures for 2010 — which was also the wettest year on record — ran 0.62 °C higher than the twentieth-century average, according to a preliminary analysis from NOAA's National Climatic Data Center in Asheville, North Carolina.

Spill science scarce

The presidential commission investigating last year's huge oil spill in the Gulf of Mexico has called for more science in federal decisions on oil production and spill response. In its final report, released last week, the commission asked Congress to supply more funding for scientific and environmental studies and to involve science agencies more formally in decisions about which areas should be opened to exploration. Commission co-chair Bob Graham said: "Science has not been given a sufficient seat at the table. Actually, I think that's a considerable understatement. It has been virtually shut out." See go.nature.com/skf5zx for more.

After artemisinin

At least US$175 million is needed to halt the spread of malaria parasites that are resistant to artemisinins, says the World Health Organization (WHO). Of this, some $60 million should be used to boost research activities including developing classes of antimalarials to replace the artemisinins, currently the most potent antimalaria drugs. Published last week, the WHO's plan for containing resistance to the drugs also calls for increased monitoring, as only 31 of the 75 countries that should be routinely testing the drugs' efficacy did so in 2010.

PEOPLE

King Faisal prize

Chemists George Whitesides, of Harvard University in Cambridge, Massachusetts, and Richard Zare, of Stanford University in California, have been announced as winners of this year's King Faisal International Prize for Science. James Thomson, of the University of Bern, and Shinya Yamanaka, of the University of California, San Francisco, and Kyoto University, Japan, took the prize for medicine for their work on stem cells. Winners receive a medal and share US$200,000 in each category.

ERC rings changes

The European Commission has appointed seven members to the governing body of the European Research Council (ERC). The appointments to the 22-member Scientific Council will begin their

2-year terms on 2 February. Fotis Kafatos, a molecular entomologist at Imperial College London who was the first president of the ERC, is among those departing. Arrivals include Nobel laureate Timothy Hunt (**pictured**), a biologist at the London Research Institute of Cancer Research UK.

BUSINESS

BP in Russian deal

BP is joining up with Russia's state oil company Rosneft to drill in the Arctic waters of the Kara Sea. The London-based company will take 9.5% of Rosneft's shares and help its Moscow-based partner explore a 125,000-square-kilometre area that it compares in potential to the North Sea. Rosneft will get 5% of BP's ordinary shares, worth around US$8 billion. Environmentalists have protested against the deal as they grow increasingly vocal about the dangers of drilling in the Arctic.

COMING UP

23–28 JANUARY
The fifth annual Arctic Frontiers conference takes place in Tromsø, Norway. Policy experts and scientists will tackle the topic of Arctic tipping points.
www.arctic-frontiers.com

24–28 JANUARY
At a workshop in Chamonix, France, scientists at CERN will discuss whether to extend the run of the Large Hadron Collider to the end of 2012 in search of the Higgs boson.
go.nature.com/udrnvx

24 JAN–4 FEB
The ninth session of the United Nations Forum on Forests — which officially kicks off the International Year of Forests, 2011 — takes place at the UN headquarters in New York.
go.nature.com/potx5i

Satellite sacking

The chief executive of a leading German space company has been suspended as from 17 January for allegedly criticizing the European satellite-navigation system Galileo, which the company is helping to build. In a statement, OHB-System, of Bremen, said it saw "no alternative" to removing Berry Smutny "in order to effectively avert any further damage to the company". Cables from the US Embassy in Berlin, obtained by WikiLeaks and publicized in the Norwegian newspaper *Aftenposten*, claim that Smutny said the Galileo project is a "stupid idea" and a "waste of money". Smutny denies saying this.

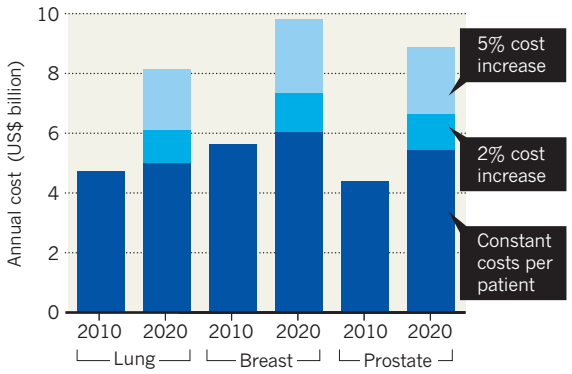
NATURE.COM
For daily news updates see:
www.nature.com/news

TREND WATCH

Even without the more expensive treatments for cancer that are to be adopted soon, the cost of caring for those with the disease will rapidly increase in the coming years. A team from the US National Cancer Institute in Bethesda, Maryland, modelled predicted changes to US population, cancer incidence and survival rates for the initial, final and continuing care stages of the disease and found that spending could rise by more than 20% by 2020. Even a small cost increase on top of that will drive numbers much higher.

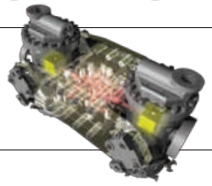
CANCER COSTS WILL KEEP CLIMBING

A 2% increase per year in the cost of initial and end-of-life care could drive total cancer costs in the United States to US$173 billion by 2020, a 39% increase. Initial care costs for three cancers are shown here.



SOURCE: NCI

NEWS IN FOCUS



S. PLATT/GETTY

About 200,000 people in Haiti have been sickened by cholera since the outbreak began in October.

PUBLIC HEALTH

Cholera vaccine plan splits experts

Opinion is divided over how to tackle the disease in Haiti.

BY DAVID CYRANOSKI, PORT-AU-PRINCE

Rarely heard in Haiti before October, 'cholera' is now an insult that children fling at one another in the teeming camps that still house more than a million people displaced by last January's devastating earthquake. Graffiti blames the disease on either the current administration — now in a contested election crisis — or the United Nations. The disease is as much a fixture in people's lives as the endless piles of rubble that remain uncleared a year after the quake.

Last week, as the country remembered the 230,000 people killed in the disaster, officials of international health agencies fine-tuned their recommendations for moving forwards with a large-scale cholera-vaccination programme. It is a controversial idea that, just months ago, with little vaccine available and the epidemic spreading rapidly, was shunned as impractical and probably ineffectual (see *Nature* **468,** 483–484; 2010). Now, with emergency care centres in place, at least in the most heavily populated areas, health officials can finally look ahead and think about how a vaccination programme might combat a disease that has become entrenched in the country.

However, *Nature*'s interviews with key partners in the proposed vaccination effort reveal significant disputes on how to proceed. Most experts in the international community recommend a limited pilot project that would determine whether to scale up and how to use cholera vaccines in future outbreaks elsewhere. The Haitian government, caught in a febrile political environment and fearful that those denied vaccination might feel resentful, is demanding immediate, broad coverage.

With no recent exposure to cholera, Haiti's population lacks natural immunity and the disease has spread quickly. Roughly 3,800 have died, with another 189,000 falling ill, since 21 October, when cholera was first recognized as the culprit. At the end of October, a local medical aid agency, GHESKIO, supported by the UN Children's Fund (UNICEF), proposed vaccinating children under five living in two slums that have not yet reported large outbreaks. "There are 200,000 people without any toilets. They collect it and dump it in the sea," says Jean-Claude Mubalama, UNICEF's chief of health in Haiti for the past five years. "If cholera arrives there, it will be very bad."

The Haitian ministry of health (MSPP) and the World Health Organization (WHO) rejected the proposal, pointing out that not enough vaccine was available. They also feared that vaccination would foster a false sense of security, causing people to relax sanitary measures; and that it would take resources away from treating the sick, or from vaccine drives against measles and other diseases. "The voice of reason was to focus on saving lives," says Jon Andrus, deputy director of the Pan American Health Organization (PAHO), the WHO's regional office. "I had driven around Port-au-Prince and seen dead bodies in the street."

In December, however, an expert committee convened by the WHO decided that vaccination should be tried, partly because they had located extra sources of the only WHO-approved vaccine, Dukoral, an expensive two-dose vaccine made by Crucell, based in Leiden, the Netherlands. On 13 January, the expert committee, including representatives from the WHO, the US Centers for Disease Control and Prevention, the US National Institutes of Health (NIH), UNICEF, the US National Vaccine Program Office and others, held a teleconference to fine-tune a vaccination plan that could form the basis of a more detailed WHO-coordinated campaign strategy. The committee is recommending a pilot project using the currently available 250,000–300,000 doses of Dukoral, and the creation of a stockpile of the vaccine for the future. ▶

The vaccination effort "can't be done nationwide and it won't have a major public-health impact", says Andrus, but it could reveal just how effective the vaccine would be in a mass immunization of a population already widely affected by cholera. Dukoral has not been used on such a scale before, although studies of thousands of people have shown it to be about 80% effective. The committee has not worked out where the campaign would be focused. "You can find areas where cholera is endemic, and that may give you a targeted population where it may have a larger impact," suggests Médecins Sans Frontières epidemiologist Kate Alberti.

"The bacterium won't go away. It has established itself."

The campaign could also help to reach the country's remote rural populations, which have a higher mortality rate. Although vaccine drives in Africa and elsewhere have faced resistance, Haitian people are eager to be vaccinated, says François Lacapère, a vaccine expert for PAHO/WHO in Haiti. Yet many Haitians are also sceptical of aid agencies' motives. Suggestions that foreigners accidentally introduced the disease (see 'How did the outbreak begin?') have given rise to unfounded rumours. Some people living in a camp that was once the Petionville golf course in Port-au-Prince, for example, make completely unsubstantiated claims that they have seen UN staff poisoning reservoirs in an attempt to further debilitate Haiti so that international powers can take over.

Even if the programme can win enough trust, using the world's entire stockpile of doses would still leave most Haitians without vaccine — a controversial prospect for the beleaguered government. Jean Ronald Cadet, the MSPP's vaccination programme manager, says the country is "90%" ready to go ahead with a campaign — but not on the small scale the WHO-convened expert group envisages.

Asked about the small pilot project proposed by the group, Cadet says "No way," shaking his head. He insists that Haiti would only consider starting to vaccinate with more than 1 million doses, with a goal of eventually reaching 6 million people. "It would depend on the pressure that the international community can put on manufacturers." Who would pay for the doses? "The international community," he says. "They brought us cholera, they have to take responsibility for taking care of it."

But mass vaccination of millions of people would necessitate much more vaccine production. About 1 million doses exist of another vaccine, Shanchol, which might be approved for use by the WHO by March (it is already approved for use in India). If production of both vaccines went into overdrive, Lacapère estimates that about 5 million doses could be prepared annually. This availability would be dependent on an advance-purchase decision, and with a six-month lag time to delivery.

Epidemiologist Renaud Piarroux of the University of the Mediterranean in Marseilles, France, says that if vaccination is going to be tried, it should be done on a large scale. "I think it can be helpful, but it should be given to millions of people in order to expect a notable effect," he says. But he doesn't see a large campaign as practical. "This will cost a lot and will require time to get a sufficient number of doses. I would prefer this money be used to improve water-supply networks and to reinforce sanitation activities," he adds. In an unpublished paper, Piarroux presents data on a large cholera outbreak in Darfur, Sudan, that happened just two years after a mass-vaccination programme, suggesting that any coverage might be of limited duration.

Others hope for a more aggressive approach. Matthew Waldor, an infectious-disease expert at Harvard Medical School in Boston, Massachusetts, says public-health officials should consider trying Peru-15, a live attenuated vaccine being developed by a consortium including Harvard Medical School and the NIH. Peru-15 is not yet in phase III trials, but has been proved safe and effective in thousands of patients, Waldor says.

Whatever approach is tried, one thing is certain: cholera is there to stay. It is likely that the bacteria now have a stronghold in Haiti's water, says Alberti. "Then you have a constant transmission between humans and the aquatic environment." With poor sanitation, little access to clean water and difficulties in reaching people to treat them — not least due to gang warfare in the slums — the country can expect repeated outbreaks, Alberti says.

Andrus agrees: "The bacterium won't go away. It has established itself." ■

➲ NATURE.COM
Read more about Haiti's cholera treatment centres:
go.nature.com/tyltvu

CHOLERA SOURCE

How did the outbreak begin?

In October 2010, rumours quickly spread in the Haitian press that a Nepalese United Nations peacekeeping base was to blame for bringing cholera to the country. An Associated Press report of excrement from the base being dumped in the Artibonite river fuelled the controversy. On 1 November, the US Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, reported genetic analyses suggesting the strain was from southeast Asia, not the one more commonly seen in Latin America. On 7 December, epidemiologist Renaud Piarroux of the University of the Mediterranean in Marseilles, France, concluded in a report for the French foreign ministry that the cases originated from near the Nepalese peacekeeping base. "I am certain that it started only a few days after a battalion of soldiers came from Katmandu, a city which was subjected to a cholera outbreak at that time," he says.

On 9 December, a more detailed genetic analysis, led by Matthew Waldor of Harvard Medical School in Boston, Massachusetts, upheld the CDC's conclusions (C. S. Chin *et al. N. Engl. J. Med.* **364,** 33–42; 2011). The analysis, based on samples from Haiti and strains from Latin America and Bangladesh, cannot definitively confirm a Nepalese source, says Waldor. The strain could have arrived with other south Asians, or from west Africa. "But the Nepalese hypothesis is fairly convincing," he says.

J. KATZ/AP PHOTO

The UN initially questioned whether it was important to pin down the origin of the cholera. But on 6 January, it appointed a committee of four scientists to investigate the claims. Waldor says that a definitive answer could come from sequencing about 20 strains — from Haiti, Nepal and West Africa, as well as more recent strains from Latin America and other south Asian countries — and by checking blood samples from Nepalese troops in Haiti for antibodies. "The truth is important," Waldor says, because it could help to understand the risk of cholera transmission from peacekeepers in the future.

Piarroux says that the international community's responsibility for the epidemic should fuel efforts to eliminate the disease. "We are responsible for importing cholera — we have to fight against it with an iron will." **D.C.**

ANTARCTIC RESEARCH

Race against time for raiders of the lost lake

Arguably the most exciting — and certainly the most controversial — scientific endeavour in Antarctica's history is close to a breakthrough.

BY QUIRIN SCHIERMEIER

A Russian drilling team is just metres away from reaching the water surface of Lake Vostok, the largest and deepest of the freshwater lakes hidden beneath Antarctica's massive ice sheet.

The ambitious project, launched more than 20 years ago, has been repeatedly delayed by technical glitches and funding problems (see *Nature* **464,** 472–473; 2010). But Russian researchers, who on 2 January resumed drilling at a depth of 3,650 metres, believe that just 20–40 metres or so of accretion ice — frozen lake water — now separate them from the lake's liquid surface. "We can make it this time," Valery Lukin, director of the Russian Antarctic programme, told *Nature*.

But time is short. Although the drill can advance by about 3 metres each day, the team must call a halt by 6 February, when the last aircraft of the summer research season is due to leave the Vostok research station, about 1,300 kilometres from the South Pole (see 'Drill for victory'). If they haven't reached the lake by then, they will have to wait until December to continue, Lukin says.

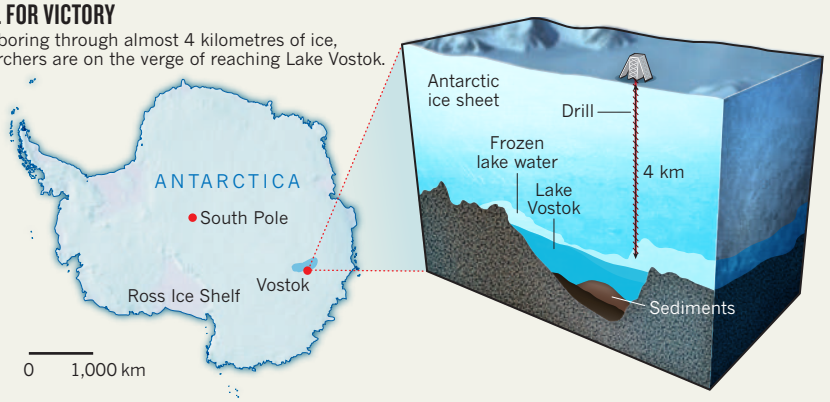
The chance of sampling one of the last uncharted environments on Earth has excited researchers ever since the lake's existence was first mooted in the 1970s. Many are thrilled by the possibility of discovering evidence of unique life forms in the lake, which is thought to have formed as much as 35 million years ago. But others worry that the drilling effort could contaminate an untouched environment. The lake may hold traces of ancient microorganisms that could reveal how life on Earth has adapted to extreme conditions.

"This is as new and exciting as flying to Mars."

At the Vostok station, tension is rising with every passing day. The team hopes that a sensor attached to the drill head will signal contact with liquid water in the next few weeks. At that point, the drill will be stopped and extracted from the bore hole, thereby lowering the pressure beneath it and drawing water into the hole. This should prevent any of the silicone drilling lubricant from entering the lake, explains Lukin. The rising water will rapidly freeze in the borehole, where drillers can extract it without penetrating the pristine lake. "If everything goes according to plan, we will re-core the hole in December and retrieve the frozen sample without polluting the lake water," he says.

DRILL FOR VICTORY

After boring through almost 4 kilometres of ice, researchers are on the verge of reaching Lake Vostok.



The plan cleared a key hurdle last November, when scientists with Russia's Arctic and Antarctic Research Institute (AARI) in St Petersburg submitted a final environmental evaluation of the project, approved by the Russian government, to the Antarctic Treaty's environmental protection committee. The document addresses queries or objections previously raised by parties to the treaty, allowing sampling operations to begin 60 days after the final evaluation was circulated to them. Scientists contacted by *Nature* acknowledge Russia's right to proceed as planned, but remain unconvinced that the sampling technology is as clean as is claimed. "From our experience there is no such thing as clean drilling," says Jean Robert Petit, a glaciologist at the Laboratory of Glaciology and Environmental Geophysics (LGGE) near Grenoble in France.

Lake Vostok has been totally isolated for almost 15 million years, and researchers suspect that it is virtually devoid of nutrients and organic carbon. Its chemistry, together with the cold, darkness and high water pressure, could mean that conditions there resemble those in the suspected ice-covered ocean on Jupiter's moon Europa. Many think that Lake Vostok's water is unlikely to support life today, but that the sediment or bedrock beneath might host microorganisms. That would feed hopes that something similar could be found on Europa.

Sediments found in accretion ice extracted in previous years by a team from the LGGE contained the thermophilic bacterium *Hydrogenophilus thermoluteolus* (S. A. Bulat *et al. Adv. Space Res.* doi:10.1016/j.asr.2010.11.024; 2010), although this does not prove that there is life in the lake itself.

If the Russians reach their goal, lake water samples will be analysed for genetic material at the AARI, where a state-of-the-art DNA sampling laboratory opened last November, says Lukin. The credibility of any findings will require meticulously documented decontamination procedures, says Martin Siegert, an Antarctic researcher at the University of Edinburgh, UK. "Scientists will rightly ask how contamination has been avoided, for example during the ice core's long journey up the borehole," he says. Lukin says that any traces of life found will be sent to foreign labs for independent verification. "We are prepared to do this properly."

Meanwhile, Russian scientists and engineers are laying plans to venture into the lake itself. In the Antarctic summer of 2012–13, they plan to send a swimming robot into the lake to collect water samples and sediments from the bottom. An environmental assessment of the plan will be submitted at the Antarctic Treaty's consultative meeting in May 2012. "We'd like to pursue this by all means," says Lukin. "For us, this is as new and exciting as flying to Mars." ■

A facility at Shidongkou No. 2 Power Plant in Shanghai, China, scrubs carbon dioxide from flue gases for a cost per tonne of CO_2 that is far below prices elsewhere.

HUANENG GROUP

CLEAN COAL

Low-cost carbon-capture project sparks interest

Consortium to determine whether price reductions seen in China can be applied abroad.

BY JEFF TOLLEFSON

The Shidongkou No. 2 Power Plant outside Shanghai, China, has hosted a parade of foreign visitors in recent months, from academics and industry officials to US energy secretary Steven Chu. All have had one question on their minds: have Chinese engineers turned a corner on carbon-capture technology?

That question occupies a small but significant place in a package of clean-energy research initiatives expected to be announced this week as Chinese President Hu Jintao meets US President Barack Obama in Washington DC from 19 January.

The buzz began in late 2009, after officials at the government-owned Huaneng Group opened a facility that captures some of the carbon dioxide emitted by the existing giant 1,320-megawatt coal-fired Shidongkou power station. The system scrubs roughly 120,000 tonnes of CO_2 a year from 3% of the facility's flue gases, but what has caught everybody's eye is the cost that Huaneng quotes: a mere US$30–35 per tonne of CO_2, including the further expense of purifying the captured gas for use in the food and beverage industry.

That is far below the $100 or more typically estimated for first-generation projects to retrofit existing power plants for carbon capture and storage (CCS) in the United States and Europe, and it is within the range of past carbon prices in the European Union emissions trading system. If similar cost reductions can be realized elsewhere, they could cut years off the timetable for commercial introduction of retrofitted CCS technology, touted as a way to reduce the climatic impact of existing coal plants. Experts want to know how the Chinese facility is doing it, and whether the savings could be exported.

During Hu's US visit this week, officials are expected to announce an initiative that will see a consortium of government, academic and business interests from the United States and China conduct a coal-technology assessment that could provide some answers. "A lot of people want to know whether that work will translate into other markets, and I believe we'll be able to shed a lot of light on that question," says Julio Friedmann, carbon-management programme leader at Lawrence Livermore National Laboratory in Livermore, California. Friedmann also serves as technical director for the US–China Advanced Coal Technology Consortium at West Virginia University in Morgantown, which was established last September as part of an energy partnership and will conduct the assessment.

➲ NATURE.COM
Read more about carbon sequestration at:
go.nature.com/dztne1

The Shidongkou retrofit builds on the work of a smaller facility, installed at the Gaobeidian power plant in Beijing in 2008 (see *Nature* **454,** 388–392; 2008). Both installations use a common CCS technology: CO_2-rich flue gas from the plant is bubbled through a column containing an amine-based solvent — in this case, a mixture of ethanolamine and additives — that reacts with the gas and takes up its carbon dioxide. The solvent is then heated to release the CO_2, and the whole process starts again. In addition to the direct costs, this process is normally expected to gulp 25% or more of a plant's energy output when fully scaled up, making it a tough sell for power companies.

Huaneng has not yet revealed all the technical details of its CCS process. Huang Bin, head of Huaneng's Research and Development Division in Beijing, says that the company has made unspecified changes in the design of the plant and the chemistry of the solvent, which increased the energy efficiency of the system by 11–14% and reduced the cost of installation by a factor of 10 per tonne of CO_2.

"Huaneng is using a known process, but they seem to have found a way to do it much more economically," says David Mohler, chief technology officer of Duke Energy, an electric utility company based in Charlotte, North Carolina. "What's the secret sauce?" he adds.

The cost that Huaneng quotes for capturing the carbon dioxide before purification — around $20 per tonne — is four to five times lower than anything anybody else is reporting, says Mohler. Duke Energy has a partnership with Huaneng and has been planning to analyse the cost of installing and running Huaneng's technology at its Gibson Station power plant in Indiana; Mohler says that joining the consortium's broader assessment will bolster the Gibson analysis and help put this technology in context with other options.

However, Howard Herzog, a chemical engineer researching carbon sequestration at the Massachusetts Institute of Technology in Cambridge, says a deeper inspection of the Shidongkou facility might reveal that its secret comes down to things such as cheap labour and fewer regulatory burdens. "The fact that it's cheaper in China doesn't impress me," says Herzog, who recently toured the facility. "Everything is cheaper in China."

Sarah Forbes, a carbon-sequestration expert at the World Resources Institute, an environmental think tank in Washington DC, says that Huaneng's costs for carbon capture are in proportion with general costs in the Chinese coal industry, which tend to be about one-third of those in the United States.

Ming Sung, who promotes US–Chinese business alliances in Beijing on behalf of the Clean Air Task Force in Boston, Massachusetts, acknowledges that costs will be higher in places such as the United States, but says that the question is how much. Assuming that Huaneng has made some industrial progress, he says, its technology will probably still be cheaper than anyone else's at this stage.

That may not be enough to jump-start the technology, Forbes cautions. "Adding carbon capture to a power plant still comes with a significant cost and energy penalty" that will discourage adoption, she says, and the regulatory drivers necessary to encourage or force take-up are still missing.

Nonetheless, EmberClear, an energy company based in Calgary, Canada, has licensed a suite of Huaneng's clean-coal technologies — including the capture technology in place in Shanghai — for deployment in the West. Albert Lin, chief executive of EmberClear, says the cost reductions that Huaneng has achieved at Shidongkou are probably the result of a mixture of practice and Chinese economics. "Half of it is Chinese costs, but there are definitely improvements in how the technology is being used," he says.

To Friedmann, Shidongkou No. 2 has become a critical case study for exploring actual CCS costs, because of its scale and the real operational data that Huaneng is providing. "There are not that many places where we are going to have the kind of data we are looking for," says Friedmann. "They are the forerunners." ■

POLICY

France mulls embryo research reform

Scientists and clinicians push for a clearer, more permissive law on human embryonic stem-cell work.

BY DECLAN BUTLER

France's tight restrictions on human embryonic stem-cell (ESC) research might soon be eased. As parliament prepares to debate the issue next month, *Nature* has learned that researchers' calls for reform are garnering political support.

Officially, research on human ESCs and embryos is banned in France. But under a 2004 amendment to the country's bioethics law, scientists can obtain dispensation for research that could lead to "major therapeutic progress" for serious diseases that resist other approaches. Those whose research fits the bill — about 30 research groups and 40 projects so far — can carry out research on whole embryos, or on cell lines derived from embryos left over from *in vitro* fertilization (IVF). Creating embryos for research purposes is illegal in the country, a position that enjoys a broad consensus among scientists, politicians and public alike.

Scientists concede that the 2004 compromise has been a great improvement on the previous outright ban. But it still left uncertainty over the regulatory status of ESCs in France. This uncertainty is a deterrent to foreign researchers and investment by companies, says Marc Peschanski, a neuroscientist working for INSERM, the national biomedical research agency, and head of the Institute for Stem Cell Therapy and Exploration of Monogenic Diseases in Evry, outside Paris. Axel Kahn, a renowned INSERM geneticist and president of the University of Paris-Descartes, calls the current law "an intellectual absurdity and a legal quirk".

A broad consensus of researchers and clinicians is now urging the government to overturn the ban, and to explicitly authorize research on ESCs and whole embryos without the need for any special dispensation.

The law is ripe for reform, says Philippe Menasché, a cardiovascular surgeon working for INSERM at the Georges Pompidou European Hospital in Paris, where his team is researching stem cells as a potential therapy for heart disorders. Conservative politicians' opposition to all forms of embryo research was so fierce in 2004 that most researchers were grateful for any progress, he recalls, but political and ideological resistance has now largely abated.

The government's draft revised bill, released last October, would maintain the existing system. But last week, a source close to the science ministry told *Nature* that the ministry will back the explicit authorization of ESC research. After five years without any apparent abuses, politicians are more comfortable with the work, he says. The successful oversight of the national Biomedicine Agency, set up in 2004 to regulate human embryology, genetics and IVF, has had a critical role in inspiring confidence, he adds.

The science ministry is holding off from taking a position on the more controversial issue of research on whole embryos, says the ministry source, preferring to wait until it has heard the views of parliament next month.

Some scientists, including Menasché and Peschanski, say they would be satisfied with a more permissive law on only ESCs, a position that has support among parliamentarians, adds Menasché. But other scientists, including Kahn, point out that ESCs are just one aspect of embryo research, and they are pushing for a more complete authorization that includes all forms of such research. This view was endorsed by the influential bipartisan Parliamentary Office for Evaluation of Scientific and Technological Options last July, and by the Conseil d'État, one of the three arms of the country's supreme court system, in a May 2009 report requested by Prime Minister François Fillon.

After five years without any apparent abuses, politicians are more comfortable with the work.

The outcome of next month's debate is "unpredictable", says Menasché. The government may decide to take the low-risk approach of maintaining the status quo, and any bill must also be approved by the highly conservative Senate. A key factor will be the recommendations of a cross-party parliamentary commission set up to examine the draft bill. Having heard testimony from researchers, it is scheduled to release its report later this month. A source says that the commission is currently "very divided". ■

STEM CELLS

Cells snag top modelling job

Heart disorder joins growing list of conditions getting the 'disease in a dish' treatment.

BY EWEN CALLAWAY

When an unconscious 28-year-old woman with a rare heart disorder was rushed to hospital, surgeons saved her life by implanting a defibrillator. In time, a sample of her cells, alive and beating in a dish, could help to save other lives.

This is the hope for patient-specific models of disease, which can be created by reprogramming the patient's cells in the lab into an undifferentiated stem-cell state and then converting them into the specialized cell types affected by the condition. A handful of such models have been created so far for diseases ranging from diabetes to rare neurodegenerative diseases (see 'Dishy models'), with many more expected soon. Pharmaceutical companies are beginning to use these cells to identify effective drug treatments and to predict side effects that may only appear in a small subset of patients.

"The patient's cells act exactly as the patient was acting when admitted to the hospital," says Lior Gepstein, a stem-cell biologist at the Technion Israel Institute of Technology in Haifa, Israel. His team created induced pluripotent stem (iPS) cells from skin cells donated by the young woman, who suffers from a genetic form of long QT syndrome. They then reprogrammed the cells into heart-muscle cells called myocardiocytes, which can be used to study the disease. Their work was published in *Nature*[1] on 16 January.

Long QT syndrome — which gets its name from a telltale anomaly in a patient's electrocardiogram — is caused by delayed electrical recharging in heart-muscle cells. The condition typically stems from inherited mutations in molecular channels that pump ions in and out of cells. People with long QT often die suddenly from arrhythmias, a kind of irregular heartbeat, in their 20s and 30s. No good animal models of the condition exist, says Gepstein, because rodent hearts beat many times faster than the human heart and use different ion channels.

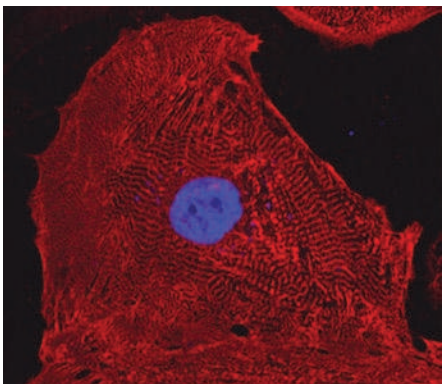
Heart-muscle cells created from the patient's iPS cells, however, recreated several aspects of her disease in a Petri dish. The cells didn't recharge as quickly as heart-muscle cells made from 'healthy' iPS cells, for example. An earlier patient-specific model of a different form of long QT syndrome found similar results[2]. Both the 'healthy' and the long QT patient-derived heart-muscle cells formed cell sheets in the Petri dish that developed a rudimentary 'heart beat'. But the patient's cells pulsed irregularly. "The normal cells of a healthy individual went boom–boom–boom; of the long QT patient they went boom–boom–boomboomboomboom," Gepstein says, doubling his tempo.

Three medicines that help heart cells recharge by targeting different kinds of ion channels all prevented arrhythmias in the diseased cells, proving the model's potential for drug screening. Moreover, the long QT patient's cells beat more irregularly when Gepstein's team treated them with a gastric-acting drug, cisapride, which was pulled from the market in 2000 because it caused lethal arrhythmia in some patients.

Predicting such side effects may be the first application for cells from long QT patients, says Gepstein. But he thinks that the eventual goal of patient-specific disease models should be personalized medicine. Clive Svendsen, director of the Cedars-Sinai Regenerative Medicine Institute in Los Angeles, agrees: "You could prescribe the best drug for a patient without rotating through different drugs and without the danger of multiple side effects."

Christine Mummery, a stem-cell biologist at the University of Leiden in the Netherlands, disagrees, saying that making personalized disease cells is too time-consuming and expensive to treat individual patients. But testing drugs against disease cells from numerous patients and comparing the cells' genetics with their responses could reveal markers that could guide treatment more generally, she says.

First, though, scientists need to overcome a shortcoming of the disease-in-the dish strategy: many patient-specific iPS cell lines don't show any obvious defects related to a disease after conversion into the relevant cell types. Gepstein's team chose long QT syndrome because it stems from genetic mutations that lead to an obvious change in the heart cells. Other conditions successfully modelled from a patient's cells, such as spinal muscular atrophy and Fanconi anaemia, also come with obvious genetic causes and clear physiological readouts. "These are all low-hanging fruits," says Mummery.



REF. 1

Heart cells derived from induced pluripotent stem cells provide a useful disease model.

Mike Venuti, president of biotech company iPierian in San Francisco, California, says that complex diseases could be modelled if the condition has a distinct effect on cells that is not present in healthy comparisons. For instance, insulin-secretion defects in patient-specific pancreatic β cells could be a proxy for diabetes. His firm has made iPS cells from people with Alzheimer's disease, Parkinson's disease and type 2 diabetes and converted them into various cell types for drug screening. He expects that drugs identified using this method will reach clinical trial for conditions such as spinal muscular atrophy in the next few years.

For all the enthusiasm about disease models made from iPS cells, embryonic stem cells (ES cells) shouldn't be forgotten for drug screens, says Stephen Minger, at GE Healthcare in Cardiff, UK. His firm is developing drug-safety tests — including for heart arrhythmias — using ES cells. He adds that the genetic manipulations used to make iPS cells could harm them and cloud the search for disease traits. "I think the field has gone a little bit overboard on iPS cells. I think they have tremendous potential, but ES cells do as well." ■

DISHY MODELS
Induced pluripotent stem cells are being used to mimic a growing number of diseases.

Condition	Cells of interest
Amyotrophic lateral sclerosis	Motor neurons[3]
Spinal muscular atrophy	Motor neurons[4]
Familial dysautonomia	Neural crest precursors[5]
Long QT	Myocardiocytes[1,2]
Fanconi anaemia	Haematopoetic progenitors[6]
Rett syndrome	Neural precursors, neurons[7]

NATURE.COM
For a video of the beating cells, see:
go.nature.com/xfwdex

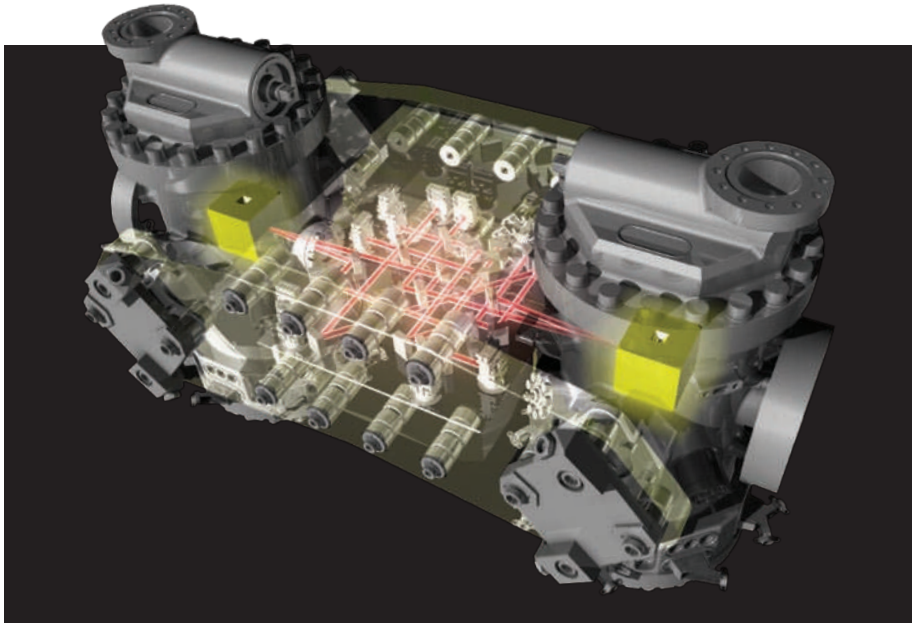
1. Itzhaki, I. *et al. Nature* doi: 10.1038/nature09747 (2011).
2. Moretti, A. *et al. N. Engl. J. Med.* **363,** 1397–1409 (2010).
3. Dimos, J. T. *et al. Science* **321,** 1218–1221 (2008).
4. Ebert, A. D. et al. *Nature* **457**, 277–280 (2009).
5. Lee, G. *et al. Nature* **461,** 402–406 (2009).
6. Raya, Á. *et al. Nature* **460,** 53–59 (2009).
7. Marchetto, M. C. N. *et al. Cell* **143,** 527–539 (2010).

ASTROPHYSICS

Missing part delays space mission

Schedule slips for European-led effort to blaze a trail for gravitational-wave detection.

BY EUGENIE SAMUEL REICH



ESA

The LISA Pathfinder is missing the mechanism to hold two masses (yellow cubes) in place during launch.

For Stefano Vitale, a principal investigator on the LISA Pathfinder mission, the situation is excruciating. Nearly all the instruments for the €300-million (US$400-million) spacecraft have been delivered for what was originally to have been a launch this year. But delays have pushed that target to 2013 and possibly later, with everything now held up by a small but crucial component. "All the rest is waiting for one part. It's heartbreaking," says Vitale, a physicist at the University of Trento in Italy.

It is a rougher-than-anticipated start for a mission that was created to find obstacles. LISA Pathfinder is a European-led test of the technology needed to run the Laser Interferometer Space Antenna (LISA), an ambitious effort to detect gravitational waves from sources in the distant Universe. Scientists hope that LISA can achieve this by measuring the precise separations between three pairs of masses free-floating inside three spacecraft positioned 5 million kilometres apart. The technical challenge along with the estimated cost of LISA (€1 billion to €2 billion) made a precursor mission a necessity. If LISA Pathfinder encounters significant problems it could sow doubts about the overall effort.

LISA Pathfinder is not expected to detect gravitational waves, but it must deploy and measure the relative positions of two test masses with sufficient precision for LISA to move forward. The missing piece of the mission is part of a 'caging mechanism' consisting of two sets of eight fingers that will hold the two 1.96-kilogram gold–platinum masses during launch, and then, once the spacecraft reaches its orbit at the L1 Lagrangian point where the gravitational pull of Earth and Sun are balanced, delicately release them. The masses will then float freely inside their separate compartments while the spacecraft uses electrical microthrusters to maintain its position so precisely that the masses do not hit the sides of their containers.

The exacting requirement for a mechanism that can hold the masses firmly enough to withstand a force of 2,000 newtons but still release them without imparting a velocity of more than 5 micrometres per second (18 millimetres an hour) lies at the heart of the delay. A first prototype of the motor powering the fingers failed key tests, prompting the European Space Agency (ESA) to set up a task force to look into the problem. The motor is now being redesigned from scratch. "Little by little, the launch date is slipping," says Pierre Binétruy of Paris Diderot University, a physicist on the LISA international science team.

Scientists on the mission say that the important thing is to learn from the delay, to avoid similar problems on LISA. With LISA Pathfinder, ESA initially followed a conventional model for managing space missions, assigning science research groups outside the space agency to design the payload — including the caging mechanism — while industrial partners designed the spacecraft itself. But designers found that the spacecraft was operationally indistinguishable from its science payload, because the positioning of the masses inside it is coupled closely to the craft's ability to keep its place in space using the microthrusters. ESA then took on the design of the caging mechanism together with a contractor, Thales Alenia Space in Milan, Italy, which was unable to comment before *Nature* went to print.

On 10 February, the ESA Science Programme Committee is expected to assess options for the new design and chart a path forwards.

Last August, the Astro2010 decadal survey of the US National Academy of Sciences ranked participation in LISA among its top priorities, above a competing project, the International X-ray Observatory (IXO). But that recommendation assumed a successful LISA Pathfinder. Xavier Barcons, a physicist at the Cantabria Institute of Physics in Santander, Spain, who works with IXO, says the problems on Pathfinder call into question the decision to rank it higher than his project. "We also have technical difficulties but we've mastered the basics. LISA is a completely new adventure," he says.

He says that it is not clear whether LISA can fly by 2025, as the decadal survey assumed. But Fabio Favata, head of ESA's science coordination office, says that by uncovering problems early, LISA Pathfinder could help LISA avoid delays. "The present situation, although unfortunate, does emphasize the importance of pathfinding," he says. ■

CORRECTIONS

The timeline in the News story 'Cancer trial errors revealed' (*Nature* **469,** 139–140; 2011) stated that Harold Varmus asked the Institute of Medicine to review Duke University's trials in June 2010. He made this request in July 2010.

The News story 'Tevatron faces final curtain' (*Nature* **469,** 141; 2011) states that the Mu2e experiment will study the decay of muons to electrons. In fact, it will look for evidence of neutrino-less conversion of muons to electrons.

The story 'Science fortunes of Balkan neighbours diverge' (*Nature* **469,** 142–143; 2011) wrongly referred to the Romanian Academy of Sciences — it should have said the Romanian Academy. Also Bulgaria cut the budget of its Academy of Sciences by 38% to 60 million leva (US$40 million), not 75 million leva as stated.

DOWN THE PETABYTE HIGHWAY

For scientists, collisions at the world's most powerful particle collider are just the start. Nature follows the torrent of data on its circuitous journey around the world.

BY GEOFF BRUMFIEL

ATLAS PARTICLE DETECTOR, SWITZERLAND, 30 MARCH 2010, *13:06* LOCAL TIME

Beneath gently rolling hills between the mountains of Switzerland and France, the world's greatest physics experiment starts its first real run. Two beams of high-energy protons meet head-on at almost the speed of light inside the Large Hadron Collider (LHC), a giant particle accelerator at CERN, Europe's high-energy physics lab. Nanoseconds after the protons crash together, their combined energy gives birth to heavier particles, which decay in an instant into a splatter of lighter debris.

At the collision point, 92 metres underground, the 7,000-tonne ATLAS detector sees everything. The debris particles pass first through the detector's inner tracker — a sophisticated layer of silicon electronics that records their paths. Beyond that lie systems that measure the energies of the particles. Some drag to a stop there, but heavy cousins of electrons called muons barrel along, flying metres from the collision point before being picked up by giant, mustard-coloured sensors.

Microprocessors convert the particles' paths and energies into electronic signals, and select a handful of promising collisions for a closer look. The data from the chosen collisions zip upstairs to a computer farm that discards the majority and creates a digital reconstruction of those that remain.

Even after rejecting 199,999 of every 200,000 collisions, the detector churns out 19 gigabytes of data in the first minute. In total, ATLAS and the three other main detectors at the LHC produced 13 petabytes (13×10^{15} bytes) of data in 2010, which would fill a stack of CDs around 14 kilometres high. That rate outstrips any other scientific effort going on today, even in data-rich fields such as genomics and climate science (see *Nature* **455,** 16–21; 2008). And the analyses are more complex too. Particle physicists must study millions of collisions at once to find the signals buried in them — information on dark matter, extra dimensions and new particles that could plug holes in current models of the Universe. Their primary quarry is the Higgs boson, a particle thought to have a central role in determining the mass of all other known particles.

The architects of the LHC decided in 2001 to deal with all that data by dividing and conquering. The results from the giant particle detectors get parcelled up and sent to a vast global network known as the Worldwide LHC Computing Grid, the most sophisticated data-taking and analysis system ever built. The network is as great a technological leap as the collider itself, and without it the project would quickly drown in its own data.

The Grid consists of some 200,000 processing cores and 150 petabytes of disk space, distributed across 34 countries through leased data lines (see 'March of the data'). By combining these resources, the Grid enables scientists to run vast analyses that would push the world's most powerful supercomputers to the edge.

CERN COMPUTING CENTRE, 30 MARCH 2010

Within minutes, the first collisions have made their way to a 1970s-era concrete building on the other side of CERN's campus. In a white, high-ceilinged room, racks containing 50,000 computing cores undertake a careful reconstruction of every selected collision. Details of each sub-detector's calibration, along with temperature readings and other environmental data from the cavern where ATLAS is housed, are used to piece each event back together. ATLAS scientists at CERN pull up reconstructions showing starbursts of narrow lines spreading from the collision points.

CERN

In Grid terminology, the CERN computing centre is known as Tier 0. It undertakes an initial analysis of the data and stores one copy. The physics data from ATLAS on the first day of the March run total about 5.2 terabytes (5.2×10^{12} bytes), enough to fill around ten laptop computers, or five of the digital storage tapes kept on the floor below the rows of processors. The first day's harvest is modest compared with what will follow, but the ATLAS experiment has more than a thousand collaborators waiting for results. If all of them logged into CERN and attempted to pull the data from the first collisions back to their home institutions, the network would grind to a halt.

So instead, the Grid automatically spreads copies of the data geographically. Inside a small partitioned section of the computing centre, a wall of panels bristles with bright-orange fibre-optic cable. This is the heart of the system, and it routes data to sites across the globe at a blistering rate of 5 gigabytes per second.

OXFORDSHIRE, UK, 30 MARCH 2010

After CERN finishes the initial analysis, a dedicated fibre-optic link carries some of the data from the first round of collisions more than 800 kilometres to the Rutherford Appleton Laboratory, a sprawling research park nestled among muddy fields in rural Oxfordshire. Here, in a modern office building, a computing farm receives the data

➲ NATURE.COM
For more on the Large Hadron Collider, visit:
go.nature.com/lxtne9

through a yellow cable only slightly thicker than a phone line. The lab is one of 11 Tier 1 centres spread around the world, where the data are further refined and split.

Particle physics is a bit like investigating a mid-air collision. Nobody is there to witness it; instead, the debris is painstakingly collected and reassembled to give investigators hints as to what happened. In this case, physicists divide up the different kinds of particles for study. One group looks at muons, for example, while another focuses on high-energy γ-rays. The computers at the lab help by creating dozens of copies of the data, focusing on various aspects of the collision. They are given names like data10_7TeV.00152166.physics_MinBias.merge.DESD_PHOJET.* — which contains data on photons and narrow jets of particles.

CHICAGO, ILLINOIS, 15 MAY 2010

A team of US researchers sends a request for data out on the Grid, and information on several subsets of the collisions from 30 March travels from Oxfordshire via New York to a post-war University of Chicago building just two blocks from the site of the Manhattan Project's first nuclear reactor.

Rob Gardner, the physicist in charge of the computing facility, says, "What we've assembled here is a data centre just about as cheaply as we can put silicon on the floor." It looks like a smaller version of the computing centres in Geneva and Oxfordshire, but with one importance difference: researchers can bring coffee into the Chicago site. "It's not a clean environment," says Gardner.

His cluster of computers is one of the Grid's 140 Tier 2 sites. Unlike Tier 1s, which undertake serious reconstructions of the data, Tier 2 centres mainly provide storage and computing resources and can be accessed by users all over the world.

In an office above the cluster, postdoc Antonio Boveia sits at a metal desk with his laptop. His machine is at the far end of the Grid from CERN, with lines of code scrolling against the black screen. To conduct an analysis — such as one on the decay of the Higgs boson into heavy particles known as W-bosons — he types in commands in the common programming language C++. For just one of Boveia's analyses, he must study tens of millions of collisions. Even if his laptop's hard drive were 4,000 times its current size and could accommodate the data, his processor would still take a few years to complete the work. "It would be impossible," he says.

The Grid makes it possible by splitting the task. When Boveia enters his request, the Grid pulls data from sites such as the one in Oxfordshire, then parcels the analysis into thousands of separate pieces and spreads it across the network. The pieces might be processed at CERN, or at a facility in Italy, or, more likely, in many places at once. In a matter of days, Boveia receives an e-mail alert telling him that the analysis is complete.

The operation does not always work so smoothly. The Tier 1 and Tier 2 centres are managed locally, which means that they each have their own protocols — and problems. In the summer of 2009, as simulated data was flowing through the Grid in advance of the first real collisions, fluff from local cottonwood trees clogged the Chicago centre's air-conditioning unit and forced a shutdown. The same year, road workers severed one of CERN's fibre-optic links in Switzerland, and a fire brought down the Tier 1 centre in Taipei, Taiwan, for months. When things go wrong, alerts are dispatched by e-mail or, occasionally, by phone to an assortment of emergency contacts around the globe.

The system relies on goodwill, says Jamie Shiers, a group leader in CERN's computing department. "We have no line management over these people whatsoever," he says. But somehow, the global cooperative produces results.

CERN, 24 DECEMBER 2010, *11:54*

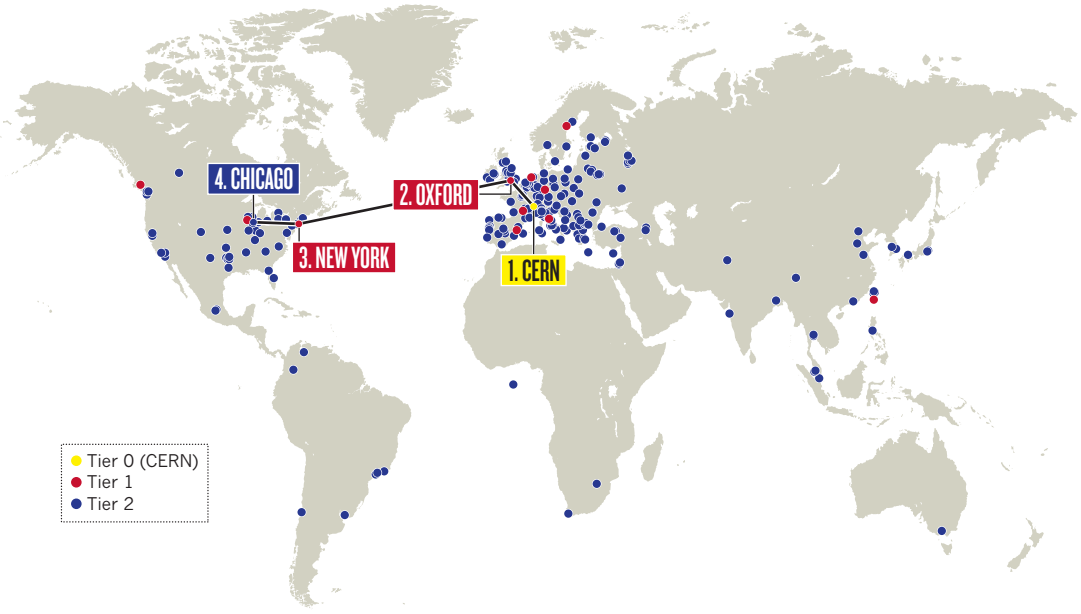
The ATLAS team posts an initial analysis from the Chicago group onto the pre-print server arXiv.org (ATLAS Collaboration. Preprint at http://arxiv.org/abs/1012.5382; 2010). The report — on W-bosons produced through mechanisms other than the decay of Higgs bosons — includes collisions from the first day's run, along with many others. Measurements of the W-bosons produced show good agreement with existing theories.

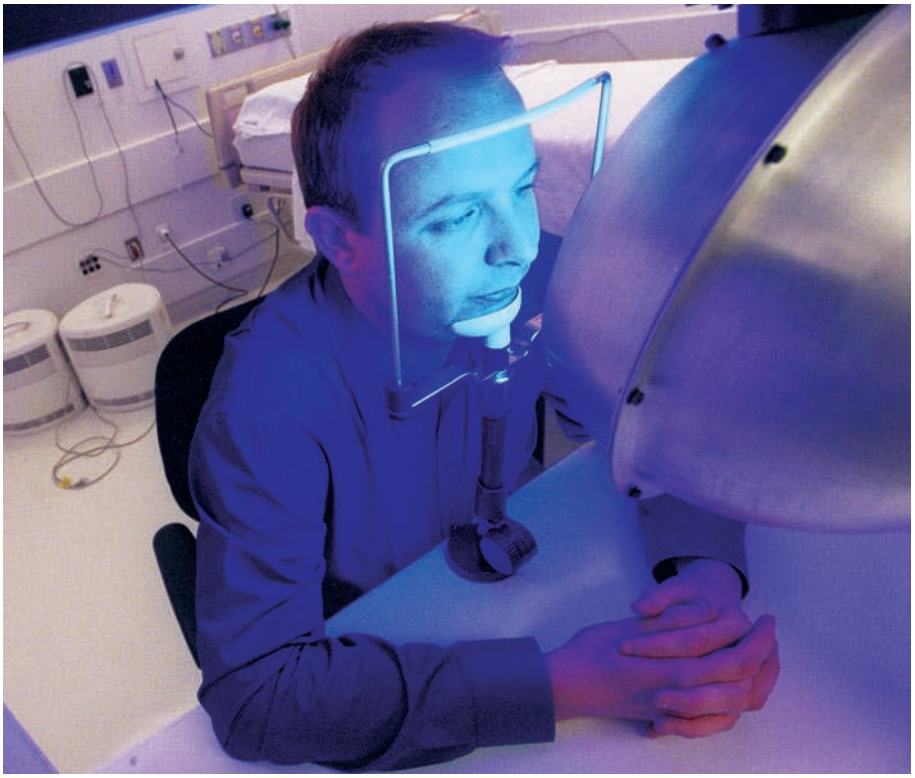
The physics data set from 30 March now makes up just 0.02% of the total data collected by the ATLAS detector. Most physicists on the collaboration are using that initial set without even realizing, as they acquire sections for analysis and combine them with other data sets. The first hints of a Higgs boson may already be stored on a computer disk in Mumbai, Melbourne or one of the many other sites to which LHC data are distributed. But even if it is there, the Higgs will stay hidden until many more petabytes have flowed through the Grid. ■

Geoff Brumfiel *is a senior reporter for* Nature *based in London.*

MARCH OF THE DATA

The Worldwide LHC Computing Grid harnesses 200,000 computing cores in 34 countries. The central node of the Grid, called Tier 0, is housed at CERN in Geneva, Switzerland; there are 11 Tier 1 sites and 140 Tier 2 sites. This story follows data on a trip from CERN to Chicago, Illinois.





K. SNIBBE/HARVARD UNIV.

Steven Lockley demonstrates an experimental set-up for studying light-sensitive cells in the eye.

SEEING WITHOUT SEEING

There is more to the eye than rods and cones — the discovery of a third photoreceptor is rewriting the visual rulebook.

BY CORIE LOK

Russell Foster remembers his first human subject, an 87-year-old woman, as she sat in a dark room facing a backlit pane of frosted glass. A genetic disorder had destroyed the light-sensing rod and cone cells in her eyes, leaving her blind for the past 50 years. She was convinced that she would see nothing. But as the wavelength of light in the room shifted to blue, she reported — after some hesitation — a sort of brightness.

"That just blew us away," says Foster, a neuroscientist at the University of Oxford, UK, and one of the senior authors of a 2007 study reporting the finding[1].

Foster and his collaborators had done nothing to treat the woman's blindness. Instead, her awareness of light owed itself to a class of light-sensitive cells discovered in 2002. Studies of these intrinsically photosensitive retinal ganglion cells (ipRGCs) have since revealed many surprises. Scientists initially thought that, rather than contribute to vision, the cells simply synchronized the circadian clock, which sets the body's 24-hour patterns of metabolism and behaviour, with changing light levels. However, recent work suggests that ipRGCs have been underestimated. They may also have a role in vision — distinguishing patterns or tracking overall brightness levels — and they seem to enable ambient light to influence cognitive processes such as learning and memory.

RODS AND CONES DETHRONED

During the past century, vision scientists focused mainly on rods and cones as the light sensors of the eye. It took Foster, an outsider coming from the circadian-biology community, to uncover some of the first evidence for a third type of photoreceptor. In the early 1990s, while at the University of Virginia in Charlottesville, his lab tested the circadian light responses in a mouse mutant with retinas that degenerate over time, and found that they were indistinguishable from the responses in mice with normal retinas. But light had no effect on the internal clocks of mice whose eyes had been removed[2].

Scepticism was strong. Foster recalls people walking out during a talk he gave. Critics of the research argued that the mutant mice probably retained some rods and cones that could be setting the clock. So in 1999, Foster, who had moved to Imperial College London, crossed transgenic mice that had no cone cells with mice that had degenerative rod cells, thus eradicating both cell types in the offspring. As long as the mice had eyes, they still had normal circadian rhythms[3,4].

The next year, Ignacio Provencio, a former graduate student of Foster's now at the University of Virginia, Charlottesville, identified the light-sensitive molecule melanopsin in the mouse and primate ganglion layer[5] — a network of retinal cells that was only thought to relay signals from rods and cones to the brain (see 'Light in layers'). The presence of this 'photopigment' suggested that some of these cells might also sense light and serve as a new class of photoreceptor. Researchers raced to isolate the cells and show that they could fire in response to light, without input from rods and cones.

The race ended in a tie in 2002. Samer Hattar, a neuroscientist at Johns Hopkins University in Baltimore, Maryland, and his colleagues found that as many as 1% of the cells in the mouse ganglion layer express melanopsin, which is most sensitive to blue light[6]. David Berson, a neuroscientist at Brown University in Providence, Rhode Island, and his lab showed that these cells, ipRGCs, detect light on their own and reach into the brain's pacemaker, the suprachiasmatic nucleus[7]. The two papers helped to win over the sceptics, says Russell Van Gelder, a neuroscientist and ophthalmologist at the University of Washington, Seattle. "Things really took off in 2002," he says.

Researchers began to develop mouse models in which they could selectively block input from each of the three photoreceptor types in the eye, to probe their individual contributions. But rather than distributing jobs neatly between cell types, the cells seem to swap roles under different conditions.

It became clear that under low light conditions, rods can set the body's clock, but some groups have suggested that under different conditions cones can as well. Perhaps more surprisingly, researchers have found that ipRGCs may contribute to visual perception. Hattar and others fluorescently labelled ipRGCs in mice to trace the projections of these cells to the brain. They found that ipRGCs reach into more brain regions than expected, including

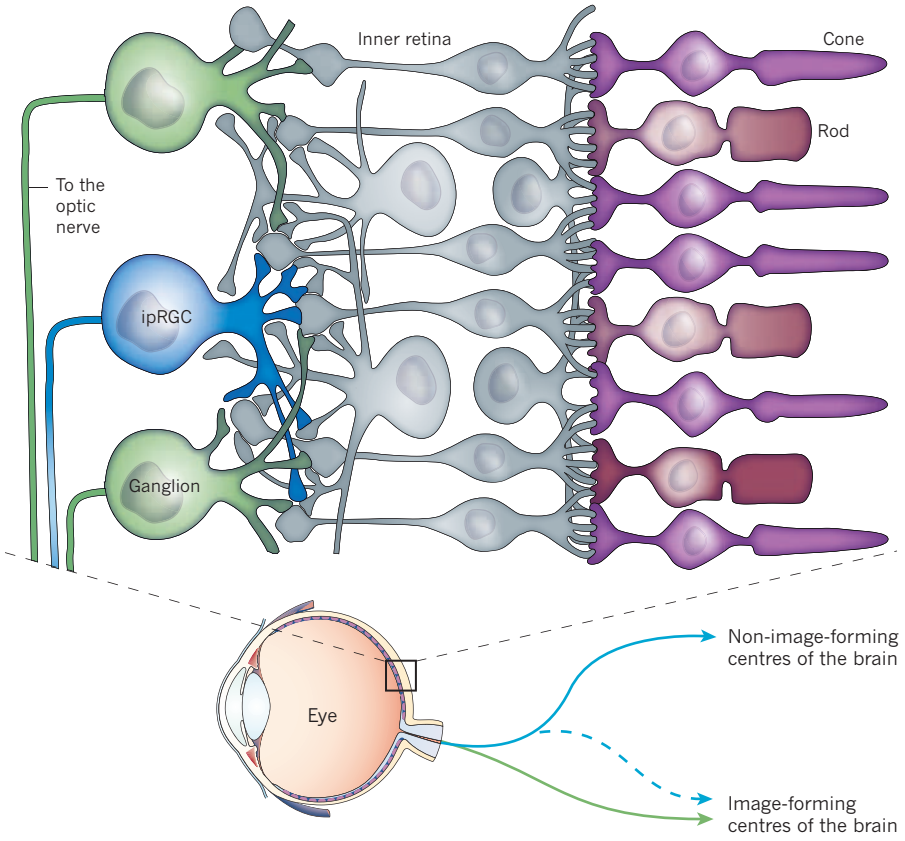
centres involved in visual processing: the dorsal lateral geniculate nucleus (LGN) and the superior colliculus. Mice without functioning rods and cones, but with intact ipRGCs, could even discriminate patterns in a visual test[8].

This is puzzling. Melanopsin responds slowly — on the order of seconds — to changes in light, limiting its ability to signal changes in spatial information, says Robert Lucas, a neurobiologist at the University of Manchester, UK. He and his group found that in mice that lacked the gene for melanopsin, and therefore had non-photoreceptive ipRGCs, almost half of the neurons in the LGN had defective light responses. The mice were unable to track background light levels, especially in the daylight range, suggesting that ipRGCs could be encoding information about brightness[9].

Researchers now think that ipRGCs and rods compensate for each other and may collectively be allowing the eyes and brain to respond to light across a wide range of brightness levels. Why these different photoreceptors share the load in such specific ways is not clear. For example, the sensitivity of ipRGCs to blue light may make them better suited to detect the arrival of dawn and dusk.

LIGHT IN LAYERS

Light passes through the ganglion layer and cells in the inner retina to the predominant photoreceptors in the eye — the rods and cones. These then send visual information back to ganglion cells, which transmit it to visual and non-visual centres of the brain. A subset of ganglion cells, called intrinsically photoreceptive retinal ganglion cells (ipRGCs), contain a photopigment, melanopsin, and can also encode and transmit information about light directly.



BEGINNING TO SEE THE LIGHT

The ipRGCs might influence phenomena beyond vision and circadian rhythms. Many physiological responses have been linked to light, such as sleep, migraine pain and seasonal affective disorder, and these have recently been associated with ipRGC activity. "There's likely to be a whole array of physiology that, to some degree, is light sensitive," says Provencio.

Learning and memory may be improved under certain light conditions. Provencio and his colleagues presented data last year showing the effects of light on a mouse model for learned fear. Mice were conditioned to associate a mild electric shock with a tone cue. Those that had learned fear in the presence of light froze for longer in response to the tone than those that had been conditioned in the dark. This effect did not appear in mice engineered to lack rods and cones, but did in melanopsin-knockout mice, suggesting that the rods and cones are driving this light-enhanced learning. Still, the researchers have not ruled out a role for ipRGCs. These cells route information from the eyes to the non-visual centres in the brain, including those involved in fear responses.

Hattar has unpublished data to suggest that activating melanopsin with light at various points in the sleep–wake cycle of mice impairs learning and memory, even when the animals have normal circadian rhythms. This could mean that exposure to light at times when the body isn't expecting it can be disruptive. And for humans, who have a smaller percentage of ipRGCs than mice, experiments are beginning to show how the cells might contribute to physiology and behaviour. Steven Lockley, a neuroscientist at Brigham and Women's Hospital in Boston, Massachusetts, and his colleagues tested the reaction times of 16 healthy volunteers while they were exposed to either blue or green light for 6.5 hours. Those exposed to blue light had faster reaction times and fewer attention lapses when they were asked to report when they heard a sound[10].

Lockley says that these different strands of research might eventually help to engineer 'healthier' light — using specific wavelengths, intensities or even patterns to activate brain pathways and improve mood, sleep or mental performance. "This research opens up a whole new field in terms of light applications, both for use therapeutically and for the general population," says Lockley.

Light of certain frequencies can have beneficial effects, but may also be detrimental to health. Lockley has been working with a group of light engineers, neuroscientists and ophthalmologists, who call themselves the Blue Light Group. They met for the first time this summer to discuss, among other things, any safety issues surrounding blue light, including the idea that excessive exposure to it might contribute to a type of vision loss known as macular degeneration. Many light-emitting diodes, a leading technology for energy-efficient lighting, are rich in blue light, points out Charles Hunt, a materials scientist at the University of California, Davis, who leads the group. Could their wider adoption lead to health problems for people?

Humans have evolved to live under natural light, says Van Gelder. "Could we be doing some damage to our health by poisoning the world with wavelengths that we're not evolved to live in?" he asks. Given the emergence of new kinds of lighting, Hunt says that it is important to find out. "We need answers quickly," he says. ■

Corie Lok *is* Nature's *Research Highlights editor.*

1. Zaidi, F. H. *et al. Curr. Biol.* **17,** 2122–2128 (2007).
2. Foster, R. G. *et al. J. Comp. Physiol. A* **169,** 39–50 (1991).
3. Freedman, M. S. *et al. Science* **284,** 502–504 (1999).
4. Lucas, R. J., Freedman, M. S, Muñoz, M., Garcia-Fernández, J. M. & Foster R. G. *Science* **284,** 505–507 (1999).
5. Provencio, I. *et al. J. Neurosci.* **20,** 600–605 (2000).
6. Hattar, S., Liao, H.-W., Takao, M., Berson, D. M. & Yau, K.-W. *Science* **295,** 1065–1070 (2002).
7. Berson, D. M., Dunn, F. A. & Takao, M. *Science* **295,** 1070–1073 (2002).
8. Ecker, J. L. *et al. Neuron* **67,** 49–60 (2010).
9. Brown, T. M. *et al. PLoS Biol.* **8,** e1000558 (2010).
10. Lockley, S. W. *et al. Sleep* **29,** 161–168 (2006).

TRIAL BY TWITTER

Blogs and tweets are ripping papers apart within days of publication, leaving researchers unsure how to react.

BY APOORVA MANDAVILLI

"Scientists discover keys to long life," proclaimed *The Wall Street Journal* headline on 1 July last year. "Who will live to be 100? Genetic test might tell," said National Public Radio a day later.

These and hundreds of similarly enthusiastic headlines were touting a paper in *Science*[1] in which researchers claimed to have identified a set of genes that could predict human longevity with 77% accuracy — a finding with potentially huge implications for medicine, health policy and the economy.

But even as the popular media was trumpeting the finding, other researchers were taking to the web to criticize the paper's methodology. "We expect that most of the results of this study will not have the same longevity as its participants," sniped a blog posted by researchers at the personal genomics company 23andMe, based in Mountain View, California.

Critics were particularly perturbed by the genome-wide association study (GWAS) that the authors had used to identify their longevity genes: the centenarians and the controls in the study had been tested with different kinds of DNA chips, which potentially skewed the results.

"Basically anybody that does a lot of GWAS knows this [pitfall], which is why we all said it so fast," says David Goldstein, director of Duke University's Center for Human Genome Variation, who voiced his concerns to a Newsweek blogger the day the study appeared.

This critical onslaught was striking — but not exceptional. Papers are increasingly being taken apart in blogs, on Twitter and on other social media within hours rather than years, and in public, rather than at small conferences or in private conversation. In December, for example, many scientists blogged immediate criticisms of another widely publicized paper[2] — this one heralding bacteria that the authors claimed use arsenic rather than phosphorus in their DNA backbone.

A CHORUS OF DISAPPROVAL

To many researchers, such rapid response is all to the good, because it weeds out sloppy work faster. "When some of these things sit around in the scientific literature for a long time, they can do damage: they can influence what people work on, they can influence whole fields," says Goldstein. This was avoided in the case of the longevity-gene paper, he says. One week after its publication, the authors released a statement saying, in part, "We have been made aware that there is a technical error in the lab test used ... [and] are now closely re-examining the analysis." Then in November, *Science* issued an 'Expression of Concern' about the paper[3], in essence questioning the validity of its results.

When asked for a comment by *Nature*, the lead investigator on the paper, Paola Sebastiani, a biostatistician at Boston University in Massachusetts, said only that she and her co-authors "feel it is premature for us to talk about our experience because this is still an ongoing issue".

For many researchers, the pace and tone of this online review can be intimidating — and can sometimes feel like an attack. How are authors supposed to respond to critiques coming from all directions? Should they even respond at all? Or should they confine their replies to the conventional, more deliberative realm of conferences and journals? "The speed of communication is ahead of the sheer time needed to think and get in the lab and work," said Felisa Wolfe-Simon, a postdoctoral fellow at the NASA Astrobiology Institute in Mountain View, California, and the lead author on the arsenic paper. Aptly enough, she circulated that comment as a tweet on Twitter, which is used by many scientists to call attention to longer articles and blog posts.

To bring some order to this chaos, it looks as though a new set of cultural norms will be needed, along with an online infrastructure to support them. The idea of open, online peer review is hardly new. Since Internet usage began to swell in the 1990s, enthusiasts have been arguing that online commenting could and

NATURE.COM
To join the debate about online review, go to:
go.nature.com/b49ej5

should replace the traditional process of pre-publication peer review that journals carry out to decide whether a paper is worth publishing.

"It makes much more sense in fact to publish everything and filter after the fact," says Cameron Neylon, a senior scientist at the Science & Technology Facilities Council, a UK funding body.

FAST FEEDBACK

In some fields, notably mathematics and physics, this sort of public discourse on a paper has long been the norm, both before and after publication. Most researchers in those fields have been depositing their draft papers in the preprint server arXiv.org for two decades. And when blogging became popular around the turn of the millennium, they were quick to start debating their research in that form.

Scientists in other fields seem less willing to get involved in pre-publication discussion. Biologists, in particular, are notoriously reluctant to publicly discuss their own work or comment on the work of others for fear of being scooped by competitors or of offending future reviewers of their own work. Adding to the disincentive is the knowledge that tenure committees and funding agencies do not explicitly reward online activity.

As a result, several journals — including, in 2005, *Nature* — have tried and mostly failed to interest scientists in various forms of open review. "Most papers sit in a wasteland of silence, attracting no attention whatsoever," says Phil Davis, a communications researcher at Cornell University in Ithaca, New York, and executive editor of The Scholarly Kitchen, a blog run by the Society for Scholarly Publishing in Wheat Ridge, Colorado.

Journals have had a little more success with post-publication peer review in the form of comments to the online versions of their papers. But the discussion is hardly vigorous, largely because the journals have usually solicited these post-publication critiques on their own websites, rather than on popular social networking sites.

"Who in their right mind is going to log on to the *PLoS One* site solely to comment on a paper?" asks Jonathan Eisen, academic editor-in-chief of *PLoS Biology*, and a prolific blogger and tweeter. "I guarantee that there are more comments on Twitter about a *PLoS* paper."

The question for researchers is how to deal with this ad-hoc analysis of papers. Unstructured, unruly and often anonymous, online commenting can be exasperating for biologists used to more conventional means of discussion. Like Sebastiani, for example, Wolfe-Simon initially tried to stay out of the brouhaha over the arsenic paper. "Any discourse will have to be peer reviewed in the same manner as our paper was, and go through a vetting process so that all discussion is properly moderated," she said when the controversy first erupted. She and a co-author later did provide answers to a few of the criticisms on her website.

But Goldstein, who has also had publications on the receiving end of negative online reviews, tries to take the process in his stride. "I think if the work is solid, it holds up over time and this chatter is not going to hurt solid work," he says. Nonetheless, he adds, "there can be a herd mentality to this, which one wants to be really careful of" — especially for examples such as the longevity and arsenic papers, for which neither the rapid spike in fame nor the equally sharp fall into disrepute may be fully justified.

One solution may lie in new ways of capturing, organizing and measuring all these scattered inputs, so that they end up making a coherent contribution to science instead of just fading back into the blogosphere. Perhaps the most successful and interesting experiments of this type can be found at websites such as Faculty of 1000 (F1000) and thirdreviewer.com, and in online reference libraries such as Mendeley, CiteULike and Zotero, which allow users to bookmark and share links to online papers or other interesting sites.

F1000, which was launched in 2002 and evaluates papers from journals across biology, is among the best known of these websites. It now relies on a 'faculty' of more than 10,000 peer-nominated researchers and clinicians who select, evaluate and rate papers with a score of 6 ('recommended'), 8 ('must read') or 10 ('exceptional'). The individual scores are then combined using a formula to generate the paper's F1000 article factor. These scores, in turn, are making some appearances in tenure packages and grant applications. "It's the only one we've been using in any systematic way," says Liz Allen, who leads post-award evaluation at the Wellcome Trust in London. "It adds another dimension to the citation index."

However, critics note that F1000 rankings tend to correlate closely with traditional citations, which suggest that they add little, if any, extra value. And most papers never attract the attention of the faculty members, so that they are never ranked at all. Even one as talked-about as the longevity paper garnered only a single rating on F1000: a must-read score of 8. For comparison, the currently highest-ranked paper on the site has an aggregate score of 62, and scores of 20 or more are common.

META-TWITTER

Given the vagaries of such measures, there is a growing interest in methods that would aggregate and quantify all of the online responses and evaluations of a paper — producing what Neylon and some others are referring to as 'alt-metrics' — and compare it with more conventional metrics.

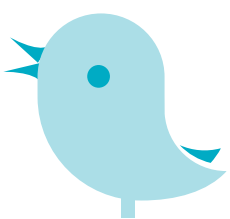
"As scholars migrate to newer forms of communication, it becomes very important to measure what they're doing and to compare," says Jason Priem, a second-year graduate student in information science at the University of North Carolina in Chapel Hill, who is focusing his study on alt-metrics.

Neylon is leading a £30,000 (US$50,000) grant proposal to create and test a working alt-metrics prototype that would rapidly measure a paper's impact by assessing all the activity surrounding it online. In addition, he and many of his colleagues champion a completely online system of pre-publication peer review that would build on the arXiv.org model, and would replace what they see as a flawed process with a more egalitarian and transparent one.

That last step, however, may be a bit farther than most scientists are willing to go — even the ones who energetically blog and tweet their post-publication reviews. Although the latter activity is "a nice secondary mechanism for catching things", says Goldstein, "I think we do not want it to be just a commentary free-for-all as the only arbiter of quality."

"It's exactly like what's said about democracy," he adds. "The peer-review process isn't very good — but there really isn't anything that's better." ■

@ironlisa The speed of communication is ahead of the sheer time needed to think and get in the lab and work.



Apoorva Mandavilli *is a writer based in New York.*

1. Sebastiani, P. *et al. Science* doi:10.1126/science.1190532 (2010).
2. Wolfe-Simon, F. *et al. Science* doi:10.1126/science.1197258 (2010).
3. Alberts, B. *Science* **330,** 912 (2010).

COMMENT

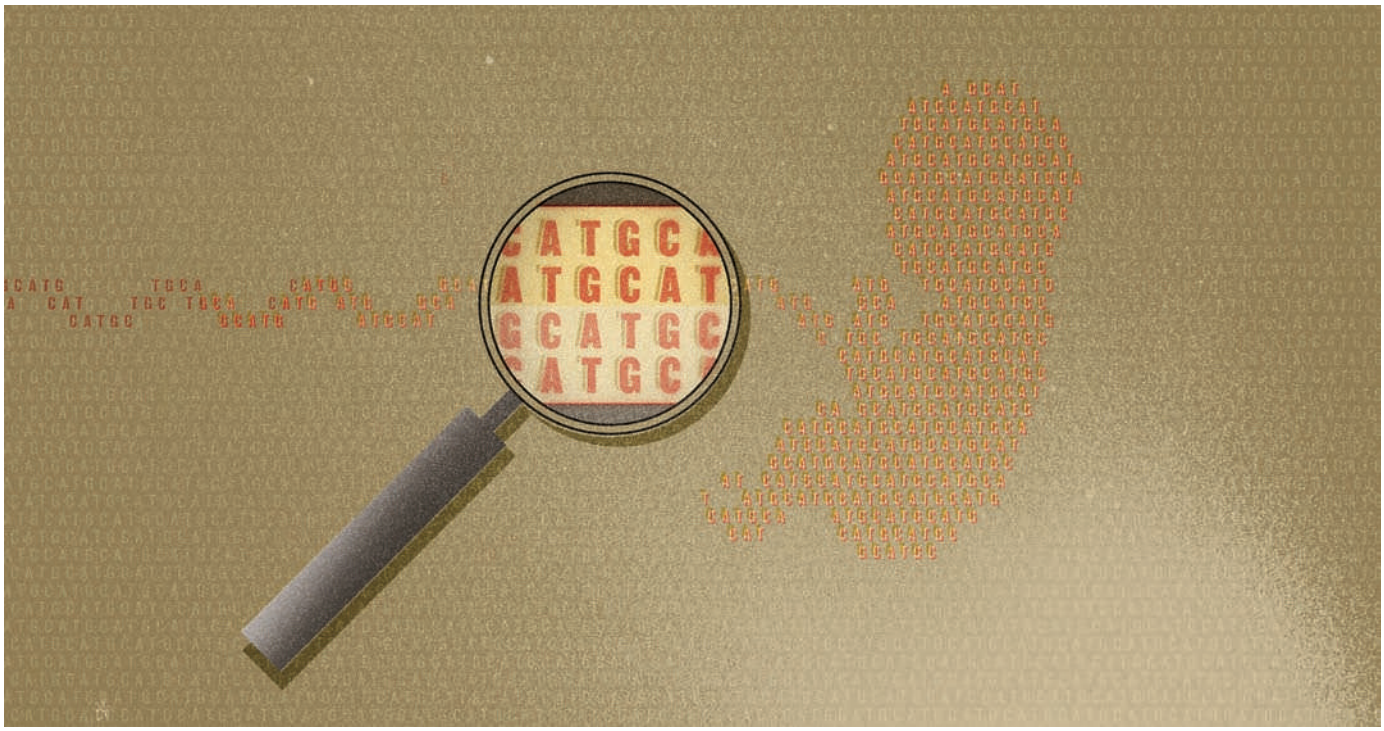


ILLUSTRATION BY GAVIN POTENZA

Get ready for the flood of fetal gene screening

Regulators, doctors and patients need to prepare for the ethical, legal and practical effects of sequencing fetal genomes from mothers' blood, says **Henry T. Greely**.

The world's news media was buzzing last week after researchers showed that a blood test for mothers could detect Down's syndrome in their fetuses[1]. Last month, two research groups independently published proof that the fetal genotype — the genetic status at a given locus — can be derived for thousands of sites from samples of fetal DNA with just a 10-millilitre blood draw from a pregnant woman[2,3].

The brave new world of widespread prenatal genetic diagnosis has been always 'arriving' since *Nature* published a paper by Danish researchers Fritz Fuchs and Povl Riis in 1956, reporting the first prenatal genetic testing in humans[4]. With non-invasive prenatal genetic diagnosis (NIPD) it may finally have arrived. Checking for hundreds or thousands of traits with one blood test, early in pregnancy, could move prenatal genetic testing from uncommon to routine.

That possibility challenges all societies to decide for which ends and by what means they want such tests to be used, raising hard questions about, among other things, abortion, disability rights, eugenics and informed consent.

Prenatal genetic testing has been clinically available since the late 1960s, but the costs, inconvenience and especially the miscarriage risks have limited its use. Each year, less than 2% of pregnant women in the United States undergo amniocentesis (in which a small amount of amniotic fluid containing fetal cells is taken for analysis) or chorionic villus sampling (CVS — in which fetal tissue is extracted from the placenta). Both procedures increase the risk of miscarriage. Until now, any given sample could be tested for only one or two conditions, typically chromosomal abnormalities such as trisomy 21, the cause of Down's syndrome.

These factors combined to limit recommended use of these methods to women with a higher risk of having a fetus with ▶

a particular disease. Most frequently this has meant women over the age of 35, whose chances of carrying a fetus with Down's syndrome are greater than the risk of a miscarriage caused by the procedure.

Biologists have known for decades that some fetal cells pass through the placenta and into the mother's blood stream. Technical problems have hampered attempts to isolate individual fetal cells and, even when such cells could be found, there was no guarantee that they were from the present pregnancy. Analysing the free-floating fragments of fetal DNA that exist in a pregnant woman's blood serum is proving more successful.

Blood contains billions of DNA fragments released when cells die and are broken up by enzymes. Even early in pregnancy, 5–10% of that 'cell-free' DNA in pregnant women comes from the fetus. Thanks to cheap and sensitive sequencing techniques, this DNA can be examined and aspects of the fetus's genome analysed. For some traits, such as paternally inherited dominant conditions, this can be done by looking for DNA variants — alleles — that the mother does not carry. For other traits, the number of copies of each variant can be used to determine how many copies of a chromosome the fetus carries, as well as how many copies of which alleles.

IN THE CLINIC

Non-invasive prenatal genetic diagnosis is already in clinical use for fetal blood-type screening. A woman of blood type Rhesus (Rh) negative can create antibodies against the red blood cells of a fetus of type Rh positive, injuring that fetus, or subsequent fetuses. In many countries, pregnant Rh-negative women routinely receive protective antibodies. Now, in several countries, including the United Kingdom, the Netherlands and France, cell-free DNA analysis is being used to determine the Rh type of the fetus and the antibody is only injected when needed.

"Commercial development of these methods seems likely within five years."

The potential of NIPD goes way beyond Rhesus screening. Two of the leading researchers in cell-free fetal DNA testing — Dennis Lo of the University of Hong Kong and Steve Quake of Stanford University in California — use different methods to analyse fetal cell-free DNA from maternal serum. Each has demonstrated the ability to detect aneuploidies — missing or extra chromosomes, such as in trisomy 21 (refs 5, 6). Last month, both researchers published proof that the fetal genotype could be derived for thousands of sites from cell-free fetal DNA[2,3] — demonstrating the possibility of using maternal blood to test for all fetal genetic traits.

There seems to be no technical barrier, given increasingly cheap genotyping and sequencing, to being able to test one sample simultaneously for chromosomal abnormalities; for single-gene diseases, such as cystic fibrosis, sickle-cell anaemia, and Tay-Sachs disease; and for various non-disease genetic traits such as sex.

Commercial firms are already interested. Sequenom in San Diego, California, is working with Lo; another, Artemis Health of Menlo Park, California, is working with Quake; and still others are also exploring the technology. For-profit development of these methods seems likely within five years, at least for chromosomal abnormalities, such as trisomy 21, and possibly for single-gene traits.

The scope and consequences of such testing will, of course, depend in large part on its accuracy. If NIPD is so inaccurate that it requires amniocentesis or CVS for confirmation, its influence will be limited. But the improving power, and decreasing cost, of DNA sequencing make it likely that the accuracy of these tests would be high. If necessary, samples can be genotyped or sequenced to greater and greater depth, particularly as costs drop, and additional samples, if needed, are just a blood draw away.

When such testing does take off — and it is when, not if — the public controversy will be about its uses. *In vitro* fertilization provides one precedent. More than 30 years from its first use, debates continue about whether it can be used by unmarried people, homosexuals and elderly women — and about who will pay for it. Preimplantation genetic diagnosis is a closer example, with strong disagreements about its use for sex selection, trait selection and the creation of 'saviour siblings'. With NIPD, abortion opponents will want little or no use of tests that will increase the number of pregnancies terminated. Some people will be concerned about technologies that prevent the birth of people with particular disabilities, both for the message that might send about the worth of those who are disabled and for its practical effects on research, treatment and support for those with disabilities.

And the spectre of eugenics will loom over the whole discussion. Some will oppose parental choices about the characteristics of their babies; others will worry that parental choice will be influenced, or trumped, by the decisions of governments, health-care systems or other institutions. Fears of eugenics will increase as such testing moves from fatal diseases to less serious medical conditions and then on to non-medical characteristics — sex selection today; skin, hair and eye colour tomorrow; perhaps, eventually, traits such as some cognitive or physical abilities. Still other kinds of uses will pose problems. Sometimes, for instance, parents with particular conditions, such as genetic forms of deafness, may want to ensure that their children have the same condition. Or some women, or the men in their lives, may want to move paternity testing *in utero*.

Some of these concerns exist today — witness for instance the dramatic skewing of live-birth sex ratios in China and India brought about by cheap and accessible ultrasound. But they will only become more immediate and more important with widespread NIPD.

Beyond these big questions lurk crucial operational details. Some involve the testing itself. Will such tests be regulated to ensure that they are safe and effective and, if so, how? Will the testing laboratories be subject to oversight that guarantees they perform the tests accurately? Who will pay for millions of genetic tests, and for the abortions that follow? The burgeoning controversies over the regulation of genetic testing, whether or not they are 'direct to consumer', provide one very contemporary example of these questions; the regulation of, and payment for, IVF and preimplantation genetic diagnosis provides another.

Much of the social impact — and the impact on the medical system — will depend on how widely such testing is used. Some of that will depend on those who fund health care and whether they see this testing as yet another cost or as a way to save money by avoiding the births of high-cost children. Part of the impact will depend on the legal system. If a test is clinically available and a physician does not offer it to a patient, at least in the United States, a physician could be liable through a 'wrongful birth' suit for the health costs of a child whose birth might have been prevented.

TIME TO TALK

In California currently, about two-thirds of pregnant women opt for non-invasive screening for Down's syndrome and neural tube defects. If the same fraction of pregnant women opt for NIPD, the United States alone would move from conducting fewer than 100,000 fetal genetic tests a year to about 3 million. Where will we find, or create, the professionals to provide genetic advice to these patients? And, of course, even if widely adopted, use of NIPD is unlikely to be uniform. It seems likely to vary between countries but also within countries, based on religion, ethnicity, education and other characteristics. In California, for example, it is thought that women with more education are more likely to accept screening and Hispanic women are less likely. What social issues will such disparities raise?

For parents who do choose NIPD, we will need to make sure they truly choose it. Today, amniocentesis and CVS are invasive procedures, typically prepared for over time. The parents and their physician decide that their fetus is at high risk of having a genetic disease, they go through genetic counselling and informed consent, and an invasive procedure

is scheduled for several days later. Confronted with a long needle or a transvaginal probe, few, if any, women will undergo either procedure without understanding that something serious is happening.

But if NIPD requires just one more tube of blood from the mother — and just one more signature on one more form — how can we ensure that parents understand what they are consenting to? Already some who get results of blood-based screening tests for the risk of Down's syndrome are shocked to learn they ever agreed to the test. NIPD greatly increases what is at stake; parents must not be surprised when genetic-test results arrive. And, of course, that consent will be even more complicated when hundreds of genetic traits can be tested, not just one or two.

"How can we ensure that parents understand what they are consenting to?"

These questions, and many others, have to be answered, and soon. Some of the answers may be the same across different cultures, others need careful national attention. Views and practices differ from country to country on abortion, on freedom of parental choice, on funding health care and on many other relevant considerations.

A few European groups have been studying NIPD. A European Union consortium called SAFE — the Special Non-invasive Advances in Fetal and Neonatal Evaluation Network — studied the scientific, medical and ethical issues around more limited NIPD applications for several years[7]; the PHG Foundation — the UK Foundation for Genomics and Population Health — convened a UK expert group that produced a report[8] on NIPD; and a more recent British project called RAPID — Reliable Accurate Prenatal non-Invasive Diagnosis — continues to work on the issues[9]. Little has been done in the United States[10] and almost nothing elsewhere.

Professional organizations, in medicine and in genetics, need to get involved, both in training their members about these technologies and in beginning to consider guidelines for their use, especially with regard to informed consent. Regulators, companies and consumer advocates need to be talking about pathways for assuring the safety, efficacy and quality of NIPD testing. In the United States, the Food and Drug Administration should start that process immediately. And it is time for ethics commissions, such as the US Presidential Commission for the Study of Bioethical Issues, to report on these issues.

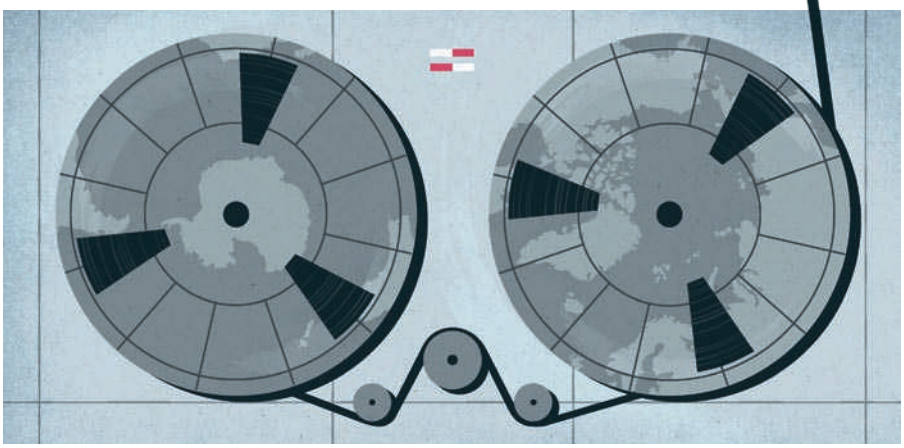
Most importantly, we need to start conversations, between all those concerned, about the limits, if any, to place on this powerful technology. Whether we view NIPD gladly as a way to reduce human suffering, warily as a step towards a eugenic dystopia, or as a mix of both, we should agree that the better we prepare, the more likely we are to avoid the worst misuses of this potentially transformative technology. ■

Henry T. Greely *is at the Center for Law and the Biosciences, Stanford Law School, Stanford, California 94305, USA.*
e-mail: hgreely@stanford.edu

1. Chiu, R. W. K. *et al. Br. Med. J.* doi: 10.1136/bmj.c7401 (2011).
2. Lo, Y. M. D. *et al. Sci. Transl. Med.* **2,** 61ra91 (2010).
3. Fan, H. C. & Quake, S. R. *Nature Precedings* doi:10.1038/npre.2010.5373.1 (2010).
4. Fuchs, F. & Riis, P. *Nature* **177,** 330 (1956).
5. Lo, Y. M. D. *et al. Nature Med.* **13,** 218–223 (2007).
6. Fan, H. C., Blumenfeld, Y. J., Chitkara, U., Hudgins, L. & Quake, S. R. *Proc. Natl Acad. Sci. USA* **105,** 16266–16271 (2008).
7. Chitty, L. S., van der Schoot, C. E., Hahn, S. & Avent, N. D. *Prenatal Diag.* **28,** 83–88 (2008).
8. Wright, C. *et al. Cell-Free Fetal Nucleic Acids for Non-Invasive Prenatal Diagnosis* (PHG Foundation, 2009).
9. Reliable Accurate Prenatal non-Invasive Diagnosis project; http://www.rapid.nhs.uk
10. The Coming Revolution in Prenatal Genetic Testing? Scientific, Ethical, Social, and Policy Responses to Maternal Serum Cell-Free Fetal DNA Testing conference (7 May 2010); audio and slides available at go.nature.com/fbhahp

A lesson in sharing

Earth scientists need better incentives, rewards and mechanisms to achieve free and open data exchange, says **David Carlson**.



JESSE LEFKOWITZ

When the polar-research community planned the International Polar Year (IPY) of 2007–08, it embraced a revolutionary goal: to establish free, open and ready access to all data. After decades of reports with 'data' and combinations of 'integrated', 'interoperable' and 'distributed' in their titles, the IPY presented an ideal test case — interdisciplinary but limited in duration and regional in focus. Yet the community found inadequate services, almost no international support and few solutions.

We have come out of the IPY with a rich burst of data, but the information uses the jargon and units of specialities from anthropology to astronomy, referenced to everything from Cartesian coordinates to postal codes. And despite the best efforts of the IPY Data and Information Service (www.ipydis.org), we cannot say how users might discover or access IPY data five years hence. Indeed, it emerged just last week that an upcoming report from the US National Academy of Sciences in Washington DC identifies the lack of data sharing as a barrier to understanding rapid changes in polar ecosystems (see *Nature* **469,** 145; 2011).

What caused these failures? Technical impediments exist relating to formats, permissions, bandwidth and so on, but the real problem is behaviour. The Earth sciences, like the science community as a whole, lack incentives for widespread data exchange.

Long before the film *Avatar* popularized it, I learned from engineering colleagues the whimsical but useful term 'unobtainium' — used to describe something perfect but elusive. A perfect data-sharing system is science's unobtainium. We must respond creatively to the challenge. Steps begun as part of the IPY by the Earth-science community include establishing a polar information commons and instigating a journal for data publication and citation.

A LEGACY OF CONFUSION

The challenges of preserving and sharing IPY data have come up repeatedly. Meteorological data from the first IPY in 1882–83 emerged in digital, accessible form only during the planning of the latest IPY. Data from the 1932–33 IPY were scattered: some were rediscovered only in recent years at the Danish Meteorological Institute in Copenhagen. By the time of the 1957–58 International Geophysical Year, the International Council for Science (ICSU) was forming World Data Centres (WDCs) to help solve the problem. There are now more than 50 WDCs, all of which pledged to support the latest IPY. Most struggled. Few received increased funding to respond to new or bigger IPY data streams, and the system had no mechanisms for handling the ecological or social threads of the IPY programme. The current WDCs, which have been supplemented by national and speciality data centres, cannot meet the needs of modern international interdisciplinary science.

The ICSU is establishing a World Data System to reform and reinvigorate the WDCs, and the World Meteorological Organization is upgrading its global information system. I endorse these efforts. But without fundamental changes to the incentives for data sharing, scientists will only perpetuate bad habits.

Data centres depend on willingness to share. All IPY projects opted in to an explicit free and open data-sharing policy (go.nature.com/byf9b4). But many researchers do not recognize, much less comply with, this policy, and few national funding organizations have the motivation or means to enforce it. Many researchers worry about others 'stealing' or misusing their data, and so hoard them.

To circumvent these attitudes, a team including myself developed the concept of a 'Polar Information Commons' (PIC; www.polarcommons.org) data label in 2009. PIC data can be freely accessed and used according to voluntary rules on attribution, citation and recognition, version control and notification, and appropriate use. The idea has been favourably received in the scientific community, and a pioneering group of polar data centres in Australia, Canada, Japan, Norway, Britain and the United States have indicated their support. However, when it comes to the nitty-gritty of making data fully available, the PIC often stalls in institutional or national legal departments. The collection of PIC-labelled data is growing — but slowly.

➲ NATURE.COM
For more on data-sharing mechanisms and practice, visit:
go.nature.com/kpfn1z

Another effort is the *Earth System Science Data* (*ESSD*) journal, which I started with Hans Pfeiffenberger, head of IT infrastructure at the Alfred Wegener Institute for Polar and Marine Research in Bremerhaven, Germany. It publishes complex and comprehensive data sets, giving data providers credit just as for a traditional publication. The journal uses a Creative Commons copyright policy to encourage free use of articles as long as the original authors and citation details are identified, and insists that authors deposit their data in a well-known open-access repository of their choice. It is a small effort so far — *ESSD* has published 28 data sets since its first issue in 2009, and remains unique in Earth sciences. We hope it will inspire similar projects.

> ***"A perfect data-sharing system is science's 'unobtainium'."***

The IPY crystallized our view of the unobtainable ideal, but hinted at solutions. The grand vision of the IPY ran aground on practicality and pragmatism, so we must take practical steps to change behaviours. ■

David Carlson *served as director of the IPY International Programme Office. He is now education and outreach director for the non-profit geodesy consortium UNAVCO in Boulder, Colorado.*
e-mail: ipy.djc@gmail.com

Izabella Godlewska de Aranda's painting *Cosmic Joy!* (2009) hints at the idea of many universes.

COSMOLOGY

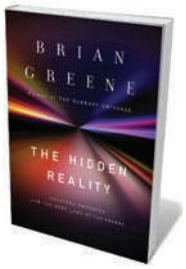
The untestable multiverse

George Ellis reminds us that Brian Greene's beguiling book on parallel worlds is more theory than fact.

Cosmology must seem odd to scientists in other fields. More and more accurate data about the distant Universe are being generated by high-tech observational techniques, giving rise to an era of 'precision cosmology' — but to explain these impressive data, cosmologists are increasingly turning to untestable theories.

In *The Hidden Reality*, theoretical physicist Brian Greene explores one of the strangest proposals: that we live in a multiverse. This fashionable concept supposes that large, perhaps infinite, numbers of separate universes exist in parallel to ours. Using well-constructed analogies, Greene explains nine different multiverse proposals, ranging from simple extrapolations of cosmological models to those based on quantum field theory, string theory and pure logic.

Greene carefully sets out the reasoning behind each proposal and the grounds for why it might be true. His arguments are absorbing and convincing on their own terms. But they inexorably lead us farther from the gold standard of testability that is the scientific method. The beguiling text moves beyond established science into philosophical speculation.



The Hidden Reality: Parallel Universes and the Deep Laws of the Cosmos
BRIAN GREENE
Knopf/Allen Lane: 2011. 384 pp. $29.95/£25

Greene's nine types of multiverse are as follows. First, if space extends forever, an infinite number of domains similar to ours might lie beyond the part of the Universe that we can see. Second, some versions of inflationary theory — the idea that the newborn Universe had a fleeting period of super-fast accelerating expansion — predict the existence of innumerable other universes, with different characteristics from our own. Third, string theory, the pre-eminent theory of quantum gravity, suggests that our Universe might be one of many four-dimensional 'braneworlds' floating in a higher-dimensional space-time.

This option is developed further in the fourth and fifth proposals, which involve cyclic universes, or variations on physical parameters that are possible in the string-theory landscape. The sixth is a quantum mechanics idea that many worlds simultaneously exist as branches of the wave function of the Universe. The seventh suggests that the Universe is a holographic projection. The eighth states that we live in one of a set of artificial universes created as simulations on a super-advanced computer. The ninth argues that it is a philosophical necessity that every possible universe must be realized somewhere, in "the grandest of all multiverses".

By presenting this plethora of theories, Greene gives the impression that the multiverse is on a sound scientific footing, but these nine arguments are mutually exclusive. We do not know how to test which is right, if any, because we cannot make direct observations of domains beyond the observational horizon — the greatest distance that light can have travelled towards us since the Universe became transparent to radiation 300,000 years after the Big Bang. Given this lack of evidence, there is a viable tenth option: that there is no multiverse at all.

Greene cites indirect evidence to support the multiverse idea. The values of the physical parameters seem to be fine-tuned to allow life. For example, if the strength of the cosmological constant — currently causing the accelerated expansion

NATURE.COM
For Hawking on the multiverse, see:
go.nature.com/zhegpz

of the Universe — was much different, then galaxies would not exist and we would not be here to make measurements. Similarly, the strength of the strong nuclear force permits atoms, and hence humans, to exist. Such anthropic reasoning invokes multiverses where there is some likelihood that physical constants take on different values in each.

But probabilistic arguments only make sense if these parallel universes actually exist. And logic cannot prove their existence. For instance, a multiverse model may predict a likely value of the cosmological constant, but the reverse is not true. A particular measurement of the cosmological constant does not require a multiverse. Nor can the multiverse concept be disproved by any specific observationally determined value of the cosmological constant, for multiverses can accommodate any value. These arguments can only provide probabilistic consistency tests for some kinds of multiverse.

So one can motivate multiverse hypotheses as plausible, but they are not observationally or experimentally testable — and never will be. It is easy to support your favourite model over others because no one can prove you wrong — you can simply adjust its parameters to fit the latest information. If the Universe is a simulation (option eight), then anything is possible. However, the existence of a computer allowing such a simulation is not remotely feasible. Scientists are beginning to confuse science with science fiction.

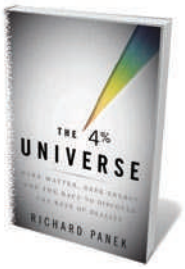
Greene, to his credit, devotes a chapter to the question of whether the multiverse idea is a scientific theory or not. He believes it is, and even supports the extravagant claim that infinities exist — infinite numbers of universes hosting countless galaxies. This leads to well-known paradoxes, such as the infinite repetition of everything because of the finiteness of possibilities. But again, there is no way to test it, because infinity is always beyond reach — and so will not plausibly exist in physical reality, as mathematician David Hilbert argued.

The gap in current theories that warrants pursuing such untestable theories is our inability to predict firmly why physical constants have the values they do. If a fundamental theory were to be proposed that explained them, the drive for a multiverse explanation would fall away. But the puzzle of why these values allow life would remain.

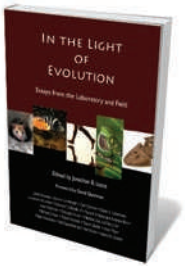
The multiverse argument is a well-founded philosophical proposal but, as it cannot be tested, it does not belong fully in the scientific fold. Read *The Hidden Reality* with enjoyment, but beware its misleading title. Greene is not presenting aspects of a known reality; he is telling of unproven theoretical possibilities. ■

George Ellis *is professor emeritus of applied mathematics at the University of Cape Town, Rondebosch 7701, South Africa.*
e-mail: george.ellis@uct.ac.za

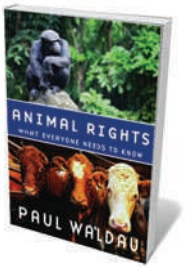
Books in brief



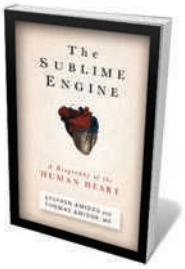
The 4% Universe: Dark Matter, Dark Energy, and the Race to Discover the Rest of Reality
Richard Panek HOUGHTON MIFFLIN HARCOURT *320 pp. $26 (2011)*
The more cosmologists have learned in recent decades, the less they seem to know about the Universe. Just 4% of it is composed of normal matter, made of protons and neutrons. The rest is 'dark matter' of unknown origin, and 'dark energy', a force that pulls space apart. Science writer Richard Panek engagingly tells the story of the discovery of these cosmic ingredients through interviews with the astronomers who unearthed them, charting the often-bitter rivalries between competing researchers.



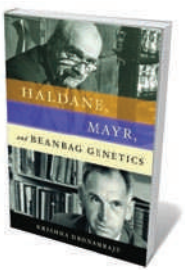
In the Light of Evolution: Essays from the Laboratory and Field
Edited by Jonathan B. Losos ROBERTS AND COMPANY PUBLISHERS *336 pp. $49.95 (2011)*
Leading experts in evolution report from the field in this collection of plainspoken essays, providing a valuable resource for non-specialists wanting to improve their understanding of this vital topic. Historian Janet Browne writes on Charles Darwin; writer Carl Zimmer muses on microbes; Daniel Lieberman discusses the evolution of human bipedalism; Marlene Zuk and Teri Orr examine sexual selection; and Neil Shubin unearths tetrapods and evolutionary steps in the fossil record.



Animal Rights: What Everyone Needs to Know
Paul Waldau OXFORD UNIVERSITY PRESS *256 pp. $16.95 (2011)*
Animal-rights theory has a long history, as legal scholar Paul Waldau describes in this primer. Setting out the basics of animal protection law, he explains our different attitudes to pets and wild animals, research and work animals, and the creatures we eat. He chronicles how our understanding of animal welfare has developed and how it has led to protective measures and legislation as well as a passionate — and often controversial — animal-rights movement. He ends with proposals for a more unified framework for animal rights.



The Sublime Engine: A Biography of the Human Heart
Stephen Amidon and Thomas Amidon RODALE BOOKS *224 pp. $24.99 (2011)*
The human heart is a fundamental organ on which we rely for our life force. For millennia, it has also been imbued with symbolism across many cultures. In their biography of the heart, writer Stephen Amidon and cardiologist Thomas Amidon trace the influence of the body's central pump through history, science, religion, literature and popular culture. From ancient Egypt and Greece, through the Middle Ages to the modern era, they ponder the miracle that is encased within our ribs.



Haldane, Mayr, and Beanbag Genetics
Krishna Dronamraju OXFORD UNIVERSITY PRESS *296 pp. $34.95 (2011)*
In the mid-twentieth century, two great biologists — J. B. S. Haldane and Ernst Mayr — clashed about the value of mathematical theories to evolution. Mayr, in 1959, queried Haldane's 'beanbag' approach to genetics, which portrayed evolution merely as "the adding of certain beans to a beanbag and the withdrawing of others". Haldane refuted Mayr's position in a witty essay in 1964. Geneticist Krishna Dronamraju relates their vigorous exchange through the scientists' correspondence.

BBC PHOTO LIBRARY

Delia Derbyshire, among other electronic-music pioneers, started out in the BBC Radiophonic Workshop.

MUSIC

Pioneers of sound

Two books chart the laboratory origins of avant-garde electronic music, finds **Marc Weidenbaum**.

Special Sound: The Creation and Legacy of the BBC Radiophonic Workshop
LOUIS NIEBUR
Oxford University Press: 2010. 272 pp. $27.95, £17.99

Begin Again: A Biography of John Cage
KENNETH SILVERMAN
Knopf: 2010. 496 pp. $40, £27.95

After decades of sonic experimentation, two maverick institutions met their ends within years of each other. In 1992, US composer John Cage, a prolific pioneer of electronic music, passed away four weeks shy of turning 80. And in 1998, the BBC Radiophonic Workshop closed down after 40 years of developing high-tech sounds. Two books — Kenneth Silverman's biography of Cage, *Begin Again*, and Louis Niebur's account of the Radiophonic Workshop, *Special Sound* — describe the technical innovations of these institutions and suggest that they were victims of their own notoriety.

Created in 1958, the BBC Radiophonic Workshop produced electronic sounds for radio and television franchises, including landmark science-fiction series such as *Quatermass*, *The Hitchhiker's Guide to the Galaxy* and, most enduringly, the signature tune for *Doctor Who*. Founded with advanced electronics equipment but a modest budget, the studio's edge was its ability to tinker within its means. It released some of its experiments to the public in pamphlets that included instructions and wiring diagrams. Eventually, electronics became so affordable that freelance composers undercut BBC economics. Alien noises, manipulated recordings and synthesized instrumentation were replicated throughout pop culture, diminishing the workshop's individuality.

The Radiophonic studio had a lab-coat reputation from the start. The small team innovated instinctively, adopting new technologies such as voltage-controlled waveform synthesis; adapting existing ones such as the turntable; and developing new sounds and approaches to enlivening narratives with audio. It owed its existence not to the BBC music department, which baulked at electronics, but to the drama and features departments, which supported the creation of a team to expand the sonic palette of their productions. The name radiophonic was chosen over another term, electrophonic, which was deemed to be too closely associated with brain research.

Born in Los Angeles in 1912, John Cage learned to tinker from his father, a serial inventor of everything from submarines to an 'Invisible Ray Vision System'. From an early age, Cage was a good orator and had a media-genic quality. Once he left home, his life took on a Zelig-like fluidity: boarding with art collector Peggy Guggenheim, befriending author John Steinbeck and studying with composer Arnold Schoenberg. In his later life, Silverman writes, Cage was so celebrated that the festivities probably contributed to his death. His birthday milestones involved exhausting global galas that curtailed his composing.

Written for ensembles large and small, many of Cage's compositions invoked chance and interactivity. He championed percussion as an instrument in Western music and developed the 'happening' (the event as art) and the 'prepared piano' (in which objects rattle on its strings). Today, Cage is best known for the silent piece less than five minutes long that contains no notes: *4'33"* was first played in 1952 and was inspired by his experience in an anechoic chamber. His technological experimentation was advanced for its time — Cage's early 1950s tape-splicing techniques, for example, did not reach their full potential until hip-hop samplers adopted them in the late 1970s and 1980s.

Many scientists were among Cage's close associates. Johan Wilhelm 'Billy' Klüver, a Bell Laboratories physicist, worked with him to invent a photoelectric system that allowed dancers to trigger sounds. Bell Labs employee Max Mathews — later of Stanford University, and after whom the widely used electronic music software Max/MSP was affectionately named — built Cage a 50-channel mixing board for Leonard Bernstein's performance of his *Atlas Eclipticalis*. And Lejaren Hiller, a former DuPont chemist, helped Cage to become computer literate in 1967, allowing him to accomplish long-envisioned projects that he had previously deemed too complex to achieve.

Begin Again — the title of which reflects the continual refreshment of Cage's ideas through his conversations with these many players — is more descriptive than critical. Silverman argues for the intensity of thought and emotion captured in Cage's body of work. Known for his philosophical musings on nature (he was an avid mycologist), as

NATURE.COM
An interview with Tod Machover, the man behind *Guitar Hero*:
go.nature.com/34hf6n

well as on chance, technology and other subjects, Cage became both a public intellectual and a celebrity.

The BBC Radiophonic Workshop also regularly replenished its perspective — by rotating its personnel. At first, this was a deliberate strategy by BBC managers. Electronic audio was so unfamiliar when the studio formed in 1958 that managers felt engineers could only work on sound effects "for a limited amount of time before succumbing to mental instability". Niebur reveals the bureaucracy and political manoeuvrings inherent in the government-funded behemoth that is the BBC. Later, many studio hands left for better-paid jobs at private companies — among them such seminal individuals in the history of electronic music as Delia Derbyshire and Daphne Oram, the latter of whom went on to influence film and opera and develop new instruments.

There are parallels between the two books. The Radiophonic faithfully served the BBC, and Cage composed reams of music for his lover Merce Cunningham's dance company. Both institutions pioneered tape loops and other techniques, and each associated closely with naturalists — Cage idolized philosopher Henry David Thoreau, and the Radiophonic team wrote ardently for wildlife broadcaster David Attenborough. Both performed for a global audience, yet their paths rarely crossed.

"Managers felt engineers could only work on sound effects 'for a limited amount of time before succumbing to mental instability'."

Differences emerge in their philosophies. Cage was intellectually highbrow, but the Radiophonic members were chartered populists. Cage loomed large in European universities and cultural institutions in Germany, France and Italy. The Radiophonic, although equally inventive, defined itself by remaining opposed to such academic forces. It avoided ideological battles about the political nature of the orchestra or about musical structures such as 12-tone serialism, in which Cage participated. The Radiophonic was, Niebur says, "the first studio in the world to make electronic music accessible to ordinary people". Its legacy lives on in the latest return to our screens of *Doctor Who* — featuring yet another rendition of that indelible theme tune. ■

Marc Weidenbaum *is a writer based in San Francisco, California, and blogs at http://disquiet.com.*
e-mail: marc@disquiet.com

ETHICS

The good life

Pascal Boyer assesses what science has to say about morals.

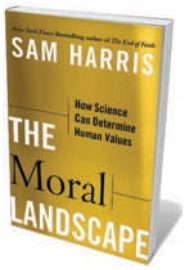
Philosopher David Hume wrote in the eighteenth century that one cannot derive an 'ought' from an 'is'. In *The Moral Landscape*, journalist Sam Harris counters this view, arguing that science can shed light on why we hold moral values and also say which values are valid. In doing so, he eloquently counters the jaded pessimists who think that science has little to say about happiness. His thesis is compelling, but he underplays the extent to which our decisions are rooted in intuition, preferring to portray decision-making as a calculated maximization of our well-being.

The psychology of morality is a rapidly growing field. Investigations of brain processes involved in moral intuitions, feelings and judgements reveal a suite of sophisticated mental capacities, fine-tuned by natural selection, that orchestrate our reactions to other people's behaviours and to our own. In distributing resources, for example, people are sensitive to differences in need and merit as well as welfare. In most circumstances, people experience moral judgements as gut feelings, produced by largely unconscious processes. Our moral reasoning is thus an awkward attempt to rationalize these intuitive feelings.

This nuanced scientific perspective is far from the hackneyed picture of virtue being a culturally derived imprint, imposed for the benefit of society on an instinctive beast. It shows how moral sentiment may have played a part in our evolution, conferring fitness advantages on ethically inclined individuals. But Harris goes further, arguing that humans can intuitively distinguish 'the good life' — which maximizes individual well-being — from less-optimal alternatives. Moral courses of action, he suggests, are those that lead us closer to this positive ideal. Scientific research addresses morality by explaining how our actions contribute to our well-being.

Harris runs against the intellectual grain in stating that moral choices can be set on an objective scale — those that increase well-being versus those that do not — and that societies are capable of moral progress, such as outlawing slavery or torture. Both views oppose the prevailing position of moral relativism, which supposes that the suffering is justifiable if it fits some local custom. Harris punctures such incoherent philosophy easily.

If morals are fluid, what should be our yardstick of valid choices? Religion is ruled out. Harris's brand of consequentialism — the ends justify the means, so what is good is what maximizes well-being — excludes transcendent sources. Because there is a quantifiable outcome, we do not need a deity to know whether our choice had good or bad consequences. Harris points out that most of the choices that are spread in the name of religion, such as proscribed killings, stem from intuition and not from the doctrine followed. Moral intuition and scientific findings are the main bases for making moral decisions.



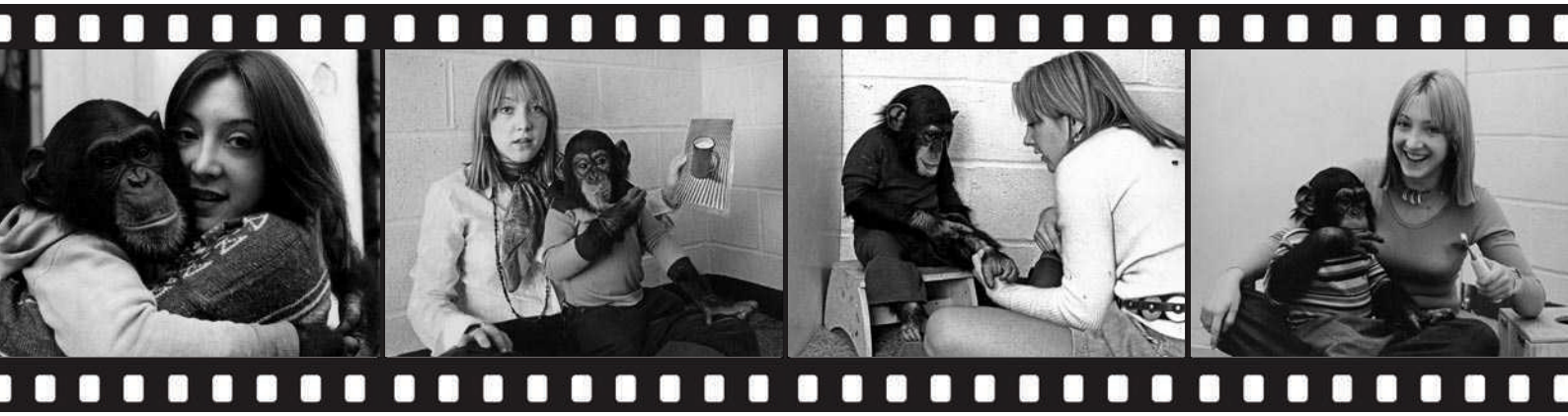
The Moral Landscape: How Science Can Determine Human Values
SAM HARRIS
The Free Press: 2010. 304 pp. $26.99

But science and intuition are not always reliable. The case against mugging people is easy to make, by quantifying suffering and benefit and the net result on the level of well-being. But an issue such as abortion is more difficult: our feelings are grounded in our intuition about whether a fetus is a person. Our idea of personhood has evolved as an approximation that is sufficient in our social interactions. Such ideas typically founder on limiting cases — we simply cannot know whether a brain-dead individual or an unborn fetus is a person or not. Harris sees such conundrums as difficult, unsolved scientific puzzles that we should pursue.

A moral optimist, Harris suggests that people can be persuaded to abandon harmful behaviours, such as the stoning of adulterers. Here, social scientists may feel that he rides roughshod over some solid findings of moral psychology. Consequentialism is not the heuristic of most humans. Experiments show that assessments of well-being are of less importance in moral decision-making than a gut feeling that actions are wrong or right. For example, beyond its genetic risks, people maintain that sibling incest is wrong, even in cases where no children result.

To be persuaded that some actions are immoral because they diminish well-being, people need to accept that welfare is the most relevant criterion of morality, which may require a special education. This and many other difficulties stand in the way of Harris's moral reforms, but they are all reasons to read his lucid, deep and uncompromising essay. ■

Pascal Boyer *is Henry Luce Professor of Individual and Collective Memory at Washington University, St Louis, USA.*



The documentary *Project Nim* charts an effort by Laura-Ann Petitto and others in the 1970s to teach sign language to a chimpanzee.

COURTESY L.-A. PETITTO

FILM

Science shines at Sundance

Relationships and behaviour are highlighted in this year's clutch of science films at the agenda-setting festival, notes **Jascha Hoffman**.

The Sundance Film Festival is known for giving Quentin Tarantino his big break as a director, and for discovering low-budget box-office hits such as *The Blair Witch Project* (1999). This year, the stars will share the screen with a Charles Darwin impersonator and an expressive chimpanzee.

The annual film festival in Park City, Utah, which this year runs 20–30 January, sets the agenda for independent cinema worldwide. Established to promote US film-makers working outside Hollywood, its popularity has grown and it now attracts big names. As in recent years, several of them have turned their cameras on scientific themes.

One point of focus is the relationship between humans and animals. James Marsh, the Oscar-winning director of *Man on Wire* (2008), presents his new documentary *Project Nim*, which charts the training of a chimpanzee in the 1970s to use hand signals. Nim Chimpsky, as the ape was called after linguist Noam Chomsky, was thought by some to have used its own syntax. The film lays scientific controversy bare by revealing how others, including the study's leader Herbert Terrace, were more sceptical.

Animal behaviour gets a comic twist in a short film by actor Isabella Rossellini. *Animals Distract Me* tracks a day in the life of Rossellini, best known for her role in David Lynch's film *Blue Velvet* (1986), as she encounters urban beasts across New York. The film follows on from her popular 2008 series *Green Porno* and its 2010 sequel *Seduce Me*, in which she plays out the mating habits of insects and sea creatures. Further insights into evolution come from a figure playing Charles Darwin, who pops up throughout.

Sundance Film Festival
Park City, Utah.
Until 30 January.

Technology forms another theme. Ridley Scott, who directed *Bladerunner* (1982), and Kevin McDonald, director of *The Last King of Scotland* (2006), will reveal at the festival the result of their project to generate a film entirely from amateur YouTube footage. *Life in a Day* is crafted from video clips of people's lives that were gathered on 24 July 2010.

In *Connected*, award-winning film-maker Tiffany Shlain, also the founder of the Webby Awards for Internet excellence, explores the global connections that have been created between people thanks to the Internet. Against the backdrop of the death during filming of her father, Leonard Shlain — surgeon and author of *Art and Physics* (William Morrow, 1993) — she asks how texts and tweets are changing our lives and relationships.

Twenty-first-century film technology is applied to dramatic visual effect in the fantastical Polish–Swedish film *The Mill and the Cross*, starring Rutger Hauer, Charlotte Rampling and Michael York. Sophisticated digital layering and colouring techniques bring to life scenes based on an oil painting of an old master, Pieter Bruegel's 1564 *The Procession to Calvary*. The artist and other period characters appear on screen as if they have stepped off the canvas.

Scientific lives lie at the heart of other poignant films. *The Music Never Stopped*, based on a case study by neurologist Oliver Sacks, depicts a young man with amnesia who recovers some of his capacity for memory when exposed to music. In *Letters from the Big Man*, a hydrologist is faced with a dilemma about whether to report her sighting of an ape-like creature in the Pacific Northwest. And in *HERE*, a US cartographer confronts uncertainty, in both map-making and love, as he works on a satellite survey of Armenia.

Two noteworthy films, co-written by and starring newcomer Brit Marling, bring out the subtle side of science fiction. In *Another Earth*, a duplicate planet to our own appears in the sky, offering a pair of strangers the chance to rewrite history on that parallel world. In *Sound of My Voice*, a couple infiltrates a cult surrounding a woman who claims to have travelled from the future. Marling's films recall another nuanced work, *Obselidia* (2010), about a man compiling an encyclopedia of the obsolete, which won last year's Alfred P. Sloan prize for feature films about science and technology at the festival.

With such variety on show, this year's Sloan award judging panel — which includes anthropologist Helen Fisher, astrophysicist Sean Carroll and neuroscientist David Poeppel — will have to pool their experience to pick a winner. "There is no all-encompassing theme," says festival panel organizer John Nein of the science and nature films. "The strongest commonality is that they offer reflections on what it means to be human." ■

Jascha Hoffman *is a writer based in San Francisco, California.*
e-mail: jascha@jaschahoffman.com

CORRESPONDENCE

Will foreign-aid pledges materialize?

Much has been made of the funding to support clean-energy and climate-adaptation initiatives in developing countries pledged by industrialized countries at the 2010 United Nations Climate Change Conference in Cancún, Mexico (*Nature* **468,** 875–876; 2010). Yet if history is any guide, it is questionable whether these funds will materialize.

Nine years ago in Monterrey, Mexico, the same industrialized nations reaffirmed their commitment to allocate 0.7% of gross national income for official development assistance (ODA). This target was originally mandated by a 1970 UN general resolution that has never been honoured. The ODA peaked in 1982, when it reached just 0.36% of the combined gross national income of countries within the Organisation for Economic Cooperation and Development (OECD).

In 2009, that proportion was languishing at 0.31%. This shortfall in annual funding amounted to US$155 billion, which is substantially more than the Cancún pledge of $100 billion by 2020. If one assumes an average annual growth rate of 2% for OECD economies, by 2020 a further $68 billion of the ODA will still be 'owed' to developing countries.

Even if the Cancún pledges are realized by 2020, they will offset only half of the deficit in the long-promised ODA. Without a more generous and decisive attitude from industrialized countries and legally binding commitments backed up by sanctions, the Cancún pledge will do little to raise the hopes of the world's two billion poorest people.

Mike Hulme *University of East Anglia, UK.*
m.hulme@uea.ac.uk

Innovation: venture capital is vital too

Technology-transfer policies stemming from the Bayh–Dole Act (*Nature* **468,** 755–756; 2010) may not always help emerging markets. They risk alienating the venture capitalists needed for commercializing early-stage innovation in places where governments control patents for taxpayer-funded research.

For example, export controls tethered to national technology-transfer policy can prevent venture capitalists from selling investments abroad, hindering the growth of a domestic industry. Ironically, under-developed local capital markets discourage capital flow for early-stage technology because of poor financial returns. This diminishes late-stage funding from banks and multinationals, thereby strangling the entire innovation system.

South Africa's nascent biotechnology industry has been a victim of such policies at the hands of the Technology Innovation Agency (TIA) — the last remaining funder for commercialization. The TIA now has an annual budget of just US$60 million or so to cover health, manufacturing, agriculture, mining, information technology and industrial biotech. BioVentures, the only life-sciences fund in sub-Saharan Africa, ranks in the top tenth of funds globally. It returns roughly three times the capital invested in it and is responsible for several successful domestic start-ups. But export controls meant that it was unable to raise a follow-up fund.

Venture capitalists also act as global intermediaries in collecting the best research and development technologies. Even with sufficient capital, government-run commercialization funders can crowd out private-sector investment and provide less-effective mentoring. This is reflected in losses of around 60 cents per dollar of biotech investment by Canadian labour-sponsored funds — a major source of venture capital.

Successful innovation requires both government and venture-capital financing. Governments must be patient and strike the right balance between overall economic and domestic innovation.

Justin Chakma *University of Toronto, Canada.*
justin.chakma@utoronto.ca
Stephen M. Sammut *University of Pennsylvania and Burrill & Company, USA.*

Cutting random funding decisions

We find that for one-third of grant applications in 2009 to Australia's National Health and Medical Research Council, success is random owing to variability among peer reviewers. Increased competition for restricted research budgets means we must rectify this element of chance in selection.

A quota limiting the number of proposals per applicant would thwart researchers who have a high success rate, while improving the odds for others. Barring unsuccessful applicants for one 'cooling-off' round is another idea (*Nature* **464,** 474–475; 2010).

Simplifying the application process would reduce costs for both applicants and peer reviewers (for example, some funding agencies request superfluous information). It would help in recruiting good peer reviewers and cut the likelihood of random decisions.

Funding for projects could be retrospective, as in some UK research institutions (D. F. Ball and J. Butler *R&D Mgmt* **34,** 87–97; 2004). Researchers would complete their research before applying, and then use the award for their next project.

Nicholas Graves, Adrian G. Barnett *Queensland University of Technology, Australia.*
n.graves@qut.edu.au
Philip Clarke *The University of Sydney, Australia.*

Bioethanol's dirty footprint in Brazil

We believe that insufficient attention is paid to the social and environmental costs incurred at regional scales by biofuel production in Brazil (A. K. Duailibe *Nature* **468,** 1041; 2010).

Brazil's Alagoas state covers almost 28,000 square kilometres, roughly half of which used to be rainforest. Sugar-cane plantations have now taken over coastal regions, including flood plains.

A study commissioned by the Alagoas government reveals that just 13.1% of the state's original rainforest has survived 35 years of the sugar-cane ethanol programme. This amounts to an average loss of 3,736 hectares of rainforest per year in what was formerly one of the world's 34 biodiversity hotspots. This environmental catastrophe is already taking its toll. Heavy rainfall in the region last year led to severe floods that destroyed thousands of buildings.

Despite academic and political controversy, most people believe biofuels to be 'clean'. In fact, ethanol production leaves a dirty footprint in one of Brazil's poorest states.

Lindemberg Medeiros de Araujo, Flávia de Barros Prado Moura *Universidade Federal de Alagoas, Brazil.*
lindemberg@pq.cnpq.br

John Fenn

(1917–2010)

Chemist who enabled mass spectrometry to weigh up biology.

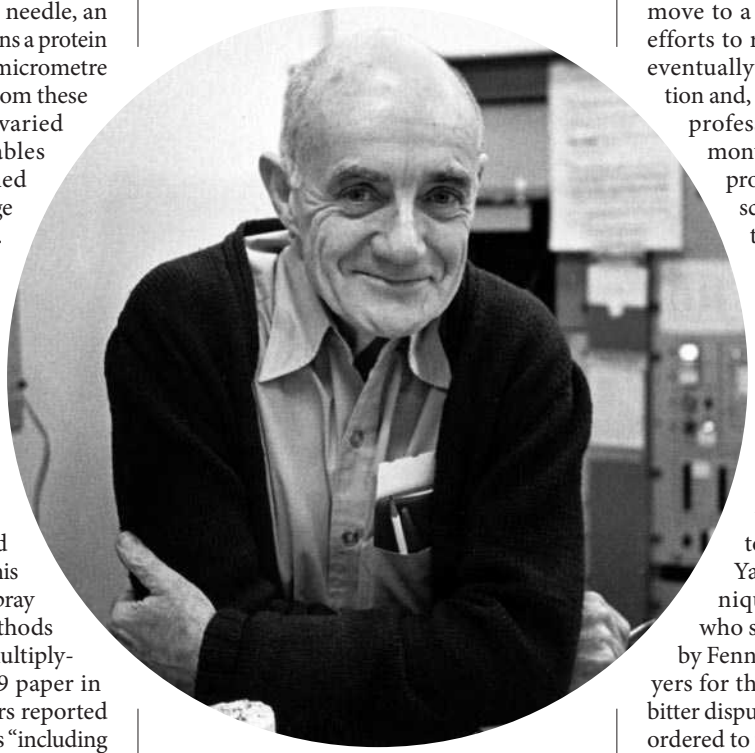
John Fenn shared the 2002 Nobel Prize in Chemistry with Koichi Tanaka and Kurt Wüthrich for his development of electrospray ionization, which, coupled with mass spectrometry, revolutionized the identification and structural analyses of large biological molecules. With characteristic wit, Fenn described the discovery as having given "wings to molecular elephants". He died on 10 December 2010, aged 93, following a fall.

Electrospray ionization enables intact proteins to transition effectively from the solution to the gas phase. By applying a high voltage (2–3 kilovolts) to a fine needle, an aerosol spray is produced that turns a protein solution into droplets in the submicrometre range (0.1–0.4 micrometres). From these fine droplets, protein ions of varied charges are formed. This enables protein masses to be determined from the series of mass-to-charge peaks in the mass spectrum. Many ions with different charges enable many measurements, so phenomenal accuracy can be achieved in the mass determined for the intact protein.

This development opened up new areas of research in the 1990s, including proteomics. For example, the molecular mass of a 17-kilodalton protein, such as myoglobin, could be measured with an error of less than 0.01%. This was not possible prior to electrospray because existing ionization methods were incapable of producing multiply-charged protein ions. In a 1989 paper in *Science*, Fenn and his co-authors reported obtaining spectra for biopolymers "including oligonucleotides and proteins, the latter having molecular weights up to 130,000, with as yet no evidence of an upper limit". This turned out to be a prophetic statement.

More than 20 years on, it is now possible to obtain well-resolved electrospray mass spectra of intact ribosomes and viruses, with molecular masses of several million daltons. At the time of the initial reports it was not clear whether non-covalent interactions could be preserved within these gas-phase protein complexes, let alone any of their three-dimensional structure. It is now clear that protein–protein interactions both survive the phase transition and retain aspects of their native structure. As a consequence, electrospray has spawned another unexpected field: gas-phase structural biology.

Fenn was born into a middle-class family in New York City in 1917, later moving with his family to Kentucky. He received his bachelor's degree in chemistry in 1937 from Berea College. The assistant registrar at Berea, Margaret Wilson, was ten years his senior; he married her in 1939. With Margaret providing the 'fellowship', in 1940 Fenn received his PhD in physical chemistry from Yale University in New Haven, Connecticut, for work on the properties of electrolyte solutions.



YALE UNIV.

After periods at industrial chemical companies, including Monsanto, Fenn joined a start-up company, Experiments, Inc., in 1945. He contributed to a large effort by the US Navy to develop a ramjet-powered anti-aircraft missile. The missile depended on pressure from high-speed flight to compress air enough to allow it to expand after combustion to provide thrust. Supersonic flight velocities were required to achieve sufficient thrust — igniting Fenn's interest in supersonic flows. Continuing this research, he moved in 1952 to Princeton University in New Jersey as director of Project SQUID, a jet-propulsion programme funded by the US Office of Naval Research.

In 1962 John returned to Yale. Here he remained until 1987, doing the majority of his Nobel-prizewinning work. In the 1960s, Fenn was interested in combustion, flame theory and stabilizing flames in high-speed flows. He decided that the way to study flame reactions was to do molecular-beam experiments. By the 1970s, he had realized that one way to get bigger molecules into the gas phase was to take a solution, disperse it, and let the solvent evaporate. So began his earliest electrospray experiments.

Meanwhile, he had to fight off a mandated move to a smaller laboratory space and efforts to retire him as he passed 70. He eventually conceded his space and position and, after a short spell as an emeritus professor, moved to Virginia Commonwealth University as a research professor in 1994. He believed that science should above all be fun and that when it ceases to be so you should give up. Fenn never gave up, continuing his research and coming into the department almost every day until just a few weeks before his death. His final paper on the mechanism of electrospray was published when he was 90.

In 2005, Fenn lost a dispute with Yale over the patent rights to electrospray. He claimed that Yale had abandoned the technique and that he was the person who saved it. A personal patent filed by Fenn was deemed 'civil theft' by lawyers for the university. After a decade of bitter dispute, and a failed appeal, Fenn was ordered to pay costs and damages totalling more than US$1 million.

Electrospray mass spectrometry continues to open doors — even in preparative mass spectrometry, where proteins and their complexes can be recovered and visualized after their flight in the gas phase. But Fenn's true legacy is the modesty and capacity to inspire that leave a lasting memory in all who had the pleasure to meet him. ■

Carol V. Robinson, *who knew John Fenn as a mentor and friend, is a Royal Society research professor at the University of Oxford Physical and Theoretical Chemistry Laboratory, Oxford OX1 3QZ, UK.*
e-mail: carol.robinson@chem.ox.ac.uk

C. GILLON/GETTY

FORUM Financial systems

Ecology and economics

A growing body of literature deals with the application of theories developed in other disciplines to financial institutions, to which a paper in this issue now adds. As outlined here, however, views differ as to its relevance. **SEE PERSPECTIVE P.351**

THE PAPER IN BRIEF

- The paper[1] is entitled 'Systemic risk in banking ecosystems' and is co-written by an expert in banking and an expert in theoretical ecology and science policy.
- It was prompted by events underlying the international financial crisis that began in 2007.
- It focuses on the network dynamics of financial institutions and, in particular, on the influence of the pricing of 'derivatives'.
- Derivatives are financial instruments that have become fiendishly complex, and that allow investment houses to hedge against, and bet on, price movements of commodities, bonds, shares and currencies without needing to hold the underlying asset.
- The authors apply models from ecology and epidemiology to explore, by a simplified 'toy model' analogy, how an initial bank failure can propagate through such institutions.
- They offer suggestions on how overall system stability can be achieved while ensuring that individual banks can maintain their necessary economic functions.

Proposing policy by analogy is risky

NEIL JOHNSON

A paper plane is a wonderful toy model with which to explain how real planes fly, and water flow is a great analogy for teaching about electrical flow through circuits. But without business-class seats, a paper plane can never be used to explain why two people pay vastly different prices for the same flight. Likewise, nobody unplugs a television to get a glass of water.

By cross-checking against our everyday experience and intuition, we can quickly see the limitations of such a toy model and analogy. However, when it comes to the complexities of the financial sector, our intuition (and arguably that of many financial experts) is so limited that rigorous statistical validation of any toy model or analogy is essential before policies are suggested. This is the potentially dangerous shortcoming of Haldane and May's paper[1]. In models of complex systems and networks, tiny changes in the model's assumptions — or changes in what it means to be a node, a link or 'infectious' — can inadvertently invert the emergent dynamics, for example by turning a stable output into an unstable one. Such changes can therefore amplify the inherent risk in any resulting policy suggestions.

There is already substantial consensus that policy-makers need to embrace financial-market risk within the framework of complex dynamic systems[2,3]. However, markets contain many heterogeneous objects, the interactions of which may change in any number of ways in the blink of an eye (or the click of a mouse). This new dynamic regime, in which the character of both the links and nodes can change on the same timescale[4,5], lies well beyond standard models of ecological food webs, disease spreading and networks. The resulting dynamic interplay can generate unexpectedly large market fluctuations — and it is these that invalidate the financial industry's existing approach to the pricing of financial derivatives and the management of risk[2,3].

The financial model[6] borrowed by Haldane and May is an interesting, abstract, complex-systems toy model. However, even the model's original creators[6] emphasized that "In order for these kind of models to be more realistic, some improvements certainly are needed". They state that their focus was on "theoretical concepts" whose "relevance for real markets requires quantitative estimates of the parameters. Given the abstract nature of the model, this appears to be a non-trivial task." They are absolutely right. Would you fly in a paper plane that had been scaled up to the size of a 747?

Policy-makers may never fully appreciate a model's limitations, so policy suggestions are potentially dangerous unless accompanied by a quantified health warning of a model's robustness and underlying assumptions, based on rigorous statistical testing against state-of-the-art financial data sets. Otherwise we simply increase risk, rather than reduce it.

Neil Johnson *is in the Department of Physics, University of Miami, Coral Gables, Florida 33146, USA.*
e-mail: njohnson@physics.miami.edu

Network theory is sorely required

THOMAS LUX

Haldane and May[1] argue that models from ecosystem research can offer valid insight for understanding the financial sector. But is this too far-fetched an analogy? Can one really imagine the regulation of financial markets being based on their similarities to networks such as food webs? My answer, in a sense, is 'yes' — we should take these similarities seriously.

This is not to say that we should equate banks, and their depositors and hedge funds, with some type of schematic predator–prey model. It is rather the potential similarities between the structural, system-wide properties of these very different areas of study that we should be interested in. As Haldane and May point out, research in biology has arrived at quite clear-cut results on the determinants of the robustness and vulnerability of ecosystems.

By contrast, the modelling of 'representative agents' in economics has led to a delusive neglect of the effects of interaction between those agents. By focusing mainly on individual optimization of utility or profit, economics has lost the perception that "more is different"[7] — namely, that higher-level aggregates (for instance, the global financial system) can have properties that cannot be understood solely on the basis of their constituent units on a lower hierarchical level (the single bank or investor). Built upon this extreme form of reductionism, the established framework for bank regulation has been exclusively micro-oriented and has lacked any consideration of systemic risk factors.

As recent history has shown, system-wide effects are important. The default of Lehman Brothers in 2008 had the contagious effects of a 'super-spreader' disease, and the subsequent domino effect brought the entire financial system to the verge of collapse. Systemic, 'macro-prudential' regulation is now an issue on the political agenda, and the structure of the financial sector has to be scrutinized for its stabilizing and destabilizing feedbacks.

However, the micro-based banking literature has little to say on such issues. We know from the natural sciences that structurally similar connections between micro units might lead to similar system behaviour in very different areas. It seems essential, therefore, to take stock of the accumulated knowledge on network structures when studying systemic risk in the banking sector. The few available phenomenological studies of particular segments of the interbank market in the 'econophysics' literature have already identified network structures that are known to be vulnerable to shocks[8,9]. The near-collapse of the overnight interbank market at various stages of the recent crisis provides an empirical confirmation of how susceptible this particular structure is to disturbances.

Connections between financial institutions are, however, multi-faceted, and only part of this complex man-made system has been mapped in existing toy models. Going beyond toy models, an empirical assessment of how trading in complex derivatives affects the network topology of the banking sector, and how that interacts with other linking factors (such as interbank credit lines), is urgently required[10]. There is, of course, also a need to go beyond the first step of analogies, and relatively simple mechanical models, to examine the behavioural micro-foundations of how the agents involved choose their connections in this financial ecosystem. ■

Thomas Lux, *currently a visiting professor at the International Christian University, Tokyo, is in the Department of Economics, University of Kiel, and the Kiel Institute for the World Economy, 24105 Kiel, Germany.*
e-mail: lux@bwl.uni-kiel.de

1. Haldane, A. G. & May, R. M. *Nature* **469,** 351–355 (2011).
2. Bouchaud, J.-P. & Potters, M. *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management* (Cambridge Univ. Press, 2009).
3. Johnson, N. F., Jefferies, P. & Hui, P. M. *Financial Market Complexity: What Physics Can Tell Us About Market Behaviour* (Oxford Univ. Press, 2003).
4. Gross, T. & Sayama, H. (eds) *Adaptive Networks* (Springer, 2009).
5. Zhao, Z. *et al. Phys. Rev. E* **81,** 056107 (2010).
6. Caccioli, F., Marsili, M. & Vivo, P. *Eur. Phys. J. B* **71,** 467–479 (2009).
7. Anderson, P. W. *Science* **177,** 393–396 (1972).
8. Soramäki, K., Bech, M. L., Arnold, J., Glass, R. J. & Beyeler, W. E. *Physica A* **379,** 317–333 (2007).
9. Iori, G., De Masi, G., Precup, O. V., Gabbi, G. & Caldarelli, G. *J. Econ. Dynam. Control* **32,** 259–278 (2008).
10. Markose, S., Giansante, S., Gatkowski, M. & Shaghaghi, A. R. *Too Interconnected To Fail: Financial Contagion and Systemic Risk in Network Model of CDS and Other Credit Enhancement Obligations of US Banks* (Univ. Essex, 2009).

NEUROSCIENCE

Seeing into the future

The resting brain recapitulates activity patterns that occurred during a recent experience, possibly to aid long-term memory formation. Surprisingly, corresponding brain activity also occurs before an event happens. SEE LETTER P.397

EDVARD I. MOSER & MAY-BRITT MOSER

Traces of experience can be found in the activity of the sleeping brain. One region in which such traces are detected is the hippocampus, which is required for episodic memory in mammals. In rodents, for example, most hippocampal neurons fire selectively when the animal is in a particular location[1,2]. When these neurons — called place cells — are active during a particular experience, they also tend to be active during subsequent sleep[3,4]. The order of firing is also often preserved[5,6]. Such subsequent replay of brain activity also occurs in the awake state, when an animal rests between bouts of running[7,8]. The recurrence of experience-related firing is thought to contribute to the reorganization of synaptic connections between neurons during memory consolidation[4,9].

However, Dragoi and Tonegawa[10] write on page 397 of this issue that hippocampal reactivation is not merely a reflection of prior experience. They initially recorded sequences of firing in place cells when mice that had been running on one arm of an L-shaped track were resting at food locations near the ends of that arm. As expected, place-cell firing sequences were replayed during pauses at the food locations. The authors then opened the other track's arm to allow the animals to run across the entire L. Surprisingly, they found that when the mice were resting before gaining access to the second arm, some of the firing sequences in their brain matched those subsequently recorded on the new arm. They refer to this predictive activity as preplay (Fig. 1).

One might argue that, on exposure of the animal to the extended segments, preplay merely reflects replay of activity associated with the familiar first part of the maze, the shape of which was similar to that of the new part. But mice with no prior experience on any track also showed preplay. Together, these observations suggest that, in a new environment, activity sequences involving a specific assembly of cells are selected from already-existing sequences in the network.

So why have other investigators using similar experimental procedures not detected preplay? After all, in the most common design, the firing sequences that occur during a behaviour are compared with resting sequences both before and after the behaviour — just as in the present study. Dragoi and Tonegawa[10] propose that their method of comparing spike sequences

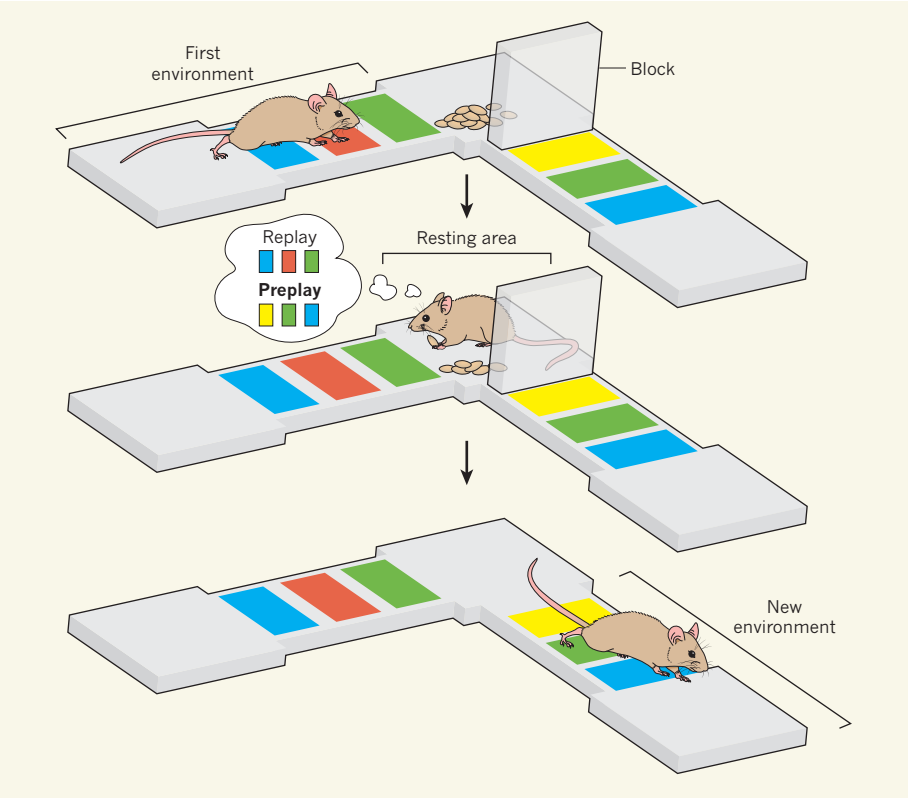


Figure 1 | Real dreams. When an animal is resting, sequences of neural activity in its brain resemble those that took place during a previous experience, suggesting that the experience is replayed. Dragoi and Tonegawa[10] show that resting mice also 'preplay' activity sequences that are predictive of subsequent activity in environments never visited before.

during rest with template sequences from the running epochs may be more sensitive than the pairwise correlation procedures used in most of the previous work. Indeed, using pairwise correlation methods, these authors[10] do not observe preplay either. It would be interesting to see whether the authors' template-based approach reveals preplay in data from past studies.

Predictive firing implies that, in the hippocampus, internal dynamics play a larger part in shaping firing sequences than has often been assumed. This deduction fits well with theoretical work[11,12] from the 1990s suggesting that the hippocampus is part of a preconfigured network for abstract representation of two-dimensional space. The metric for this internal map was proposed to be self-motion, with environment-specific information being added only secondarily — through learning.

Although self-motion-dependent signals are now thought to originate outside the hippocampus[2,13], this brain region may receive and express such signals. Consistent with this idea, place cells continue to fire in regular sequences when an animal's position is fixed, for example when a rat is running in a wheel[14]. Moreover, rat pups exploring an open space for the first time show adult-like place-cell sequences[15,16], which indicates that path sequences are hardwired in the synaptic-connection matrix by either genetic programs or early experience. As Dragoi and Tonegawa[10] show, in the adult brain such pre-wired sequences can be retrieved both during sleep and while running in a new environment.

Does preplay rule out experience as a major determinant of sequence reactivation? The answer is no. Dragoi and Tonegawa report that, although the order of firing on the track resembled firing sequences both before and after the running session, there were roughly three times as many replay events as preplay events. This observation is in agreement with the previously reported enhanced correlation after a behaviour. It seems, therefore, that although some sequences might be hardwired, others are shaped by recent experience.

The fact that replay and preplay occur at different times provides further support for mechanistic segregation. Factors that control when replay and preplay take place should be identified, as well as any interaction between these activites. For instance, are replay sequences moulded from pre-existing templates expressed during preplay? And, if so, do such preconfigured sequences contribute to generalization of experience between similar environments[17]?

Dragoi and Tonegawa[10] show that as many as 15% of the spiking events that occurred during resting correlated significantly with subsequent sequences in the new environment. This relatively large number of predictive-spiking events is worthy of some reflection. If preconfigured sequences were used as a raw material to encode sequences in new environments, and each environment received a unique set of firing sequences, then the hippocampus might soon run out of templates. It is possible, therefore, that replay and preplay originate from different neuronal pools, and that the fraction of the two event types remains constant.

One possibility is that replay depends on experience-sensitive associative networks in the CA3 area of the hippocampus, whereas preplay reflects rigid, environmentally neutral path sequences coming in from regions outside the hippocampus such as the medial entorhinal cortex[2,13]. Such hardwired entorhinal sequences could also account for the presence, in hippocampal-cell ensembles, of activity sequences corresponding to routes never taken by the animal[18]. Another possibility is that preplay reflects a general tendency for neurons to carry over firing patterns from one environment to the next when the two different environments are visited close together in time. Indeed, such carry-over effects have been observed[19] in the CA1 area of the hippocampus. Whether predictive activity disappears at intervals longer than the time constant of CA1 hysteresis should be determined. Researchers will undoubtedly address these possibilities, as well as the mechanisms and functions of preplay more generally. ■

Edvard I. Moser *and* **May-Britt Moser** *are at the Kavli Institute for Systems Neuroscience, Norwegian University of Science and Technology, Box 8905, 7491 Trondheim, Norway.*
e-mail: edvard.moser@ntnu.no

1. O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map* (Clarendon, 1978).
2. Moser, E. I., Kropff, E. & Moser, M.-B. *Annu. Rev. Neurosci.* **31,** 69–89 (2008).
3. Pavlides, C. & Winson, J. *J. Neurosci.* **9,** 2907–2918 (1989).
4. Wilson, M. A. & McNaughton, B. L. *Science* **265,** 676–679 (1994).
5. Skaggs, W. E. & McNaughton, B. L. *Science* **271,** 1870–1873 (1996).
6. Lee, A. K. & Wilson, M. A. *Neuron* **36,** 1183–1194 (2002).
7. Foster, D. J. & Wilson, M. A. *Nature* **440,** 680–683 (2006).
8. O'Neill, J., Senior, T. & Csicsvari, J. *Neuron* **49,** 143–155 (2006).
9. Buzsáki, G. *Neuroscience* **31,** 551–570 (1989).
10. Dragoi, G. & Tonegawa, S. *Nature* **469,** 397–401 (2011).
11. McNaughton, B. L. *et al. J. Exp. Biol.* **199,** 173–185 (1996).
12. Samsonovich, A. & McNaughton, B. L. *J. Neurosci.* **17,** 5900–5920 (1997).
13. Hafting, T., Fyhn, M., Molden, S., Moser, M.-B. & Moser, E. I. *Nature* **436,** 801–806 (2005).
14. Pastalkova, E., Itskov, V., Amarasingham, A. & Buzsáki, G. *Science* **321,** 1322–1327 (2008).
15. Wills, T. J., Cacucci, F., Burgess, N. & O'Keefe, J. *Science* **328,** 1573–1576 (2010).
16. Langston, R. F. *et al. Science* **328,** 1576–1580 (2010).
17. Tse, D. *et al. Science* **316,** 76–82 (2007).
18. Gupta, A. S., van der Meer, M. A. A., Toretzky, D. S. & Redish, A. D. *Neuron* **65,** 695–705 (2010).
19. Leutgeb, J. K. *et al. Neuron* **48,** 345–358 (2005).

ASTROPHYSICS

How galaxies got their black holes

The massive compact objects in the centres of galaxies developed in at least two ways. One seems to be a natural result of galaxy formation in the Big Bang theory of the expanding Universe — the other is enigmatic. SEE LETTERS P.374 & P.377

P. JAMES E. PEEBLES

A large galaxy usually has a central, compact massive object, termed a relativistic black hole for want of a better idea of what it is, that can produce great bursts of energy. When the black hole is surrounded by a cloud of old stars — the 'bulge' of the host galaxy — its mass is a few per cent of the mass of the bulge. This relationship, observed for black-hole masses ranging from about 1 million to 1,000 million solar masses, suggests that black holes and bulges evolved together. Other large, spiral-shaped galaxies have black holes of respectable mass (1 million to 100 million solar masses) and no perceptible bulge. In two papers in this issue, Kormendy *et al.*[1,2] explore clues to how these bulgeless, or pure-disk, spirals and their black holes formed.

Most of the mass in galaxies does not exist in the form of stars but in a halo of dark matter (matter different from the hydrogen and heavier elements of which we and the stars are made). There is a substantial literature on the elegant idea that dark matter controls the size of the central black hole, but Kormendy and Bender[2] (page 377) show that this cannot be so. The nearby bulgeless spiral galaxy M101 (Fig. 1) illustrates the situation: if M101 has a central black hole, its mass must be tiny relative to that of the black holes of other spirals with similar dark-matter haloes.

Bulges and their black holes seem to be a natural consequence of structure formation in the hot Big Bang theory of the expanding Universe. According to this theory, galaxies grew by gravitational assembly of matter into clumps that gathered into larger clumps, and so on to galaxies. In galaxies with bulges, including ellipticals, which have bulges and no disks, the mass of the central black hole correlates not only with the mass of the bulge, but also, as Kormendy, Bender and Cornell[1] note (page 374), with the average spread of velocities of the bulge stars (see Fig. 2a on page 375). The plausible explanation is that part of the gas out of which bulge stars formed settled instead near to the black hole, in part increasing its mass and in part fuelling explosions that blew the gas away and suppressed bulge-star formation. That is, the growth of bulge and black hole may have controlled each other. The timing looks right. Bulge stars are old: they formed when the expanding Universe was roughly a third of its present size (redshift about 2). This is when the rate of star formation per unit of matter was near its maximum (more than 10 times the present rate[3]). It is also when quasars — explosions powered by the central black holes — were most abundant (100 times more common than now[4]), probably an explosive result of overfeeding of the black holes as the early generations of stars were forming.

In addition to the evidence that black holes and bulges co-evolve, there are indications of the possibly related evolution of other components of galaxies. The mass distribution in a spiral galaxy varies smoothly from its outer, dark-matter-dominated parts to its inner parts, where the mass in stars is important. That is, the dark matter seems to have had a strong influence on the formation of the spiral galaxy's disk of gas, stars and dust — but not on the formation of the central black hole.

Another plausible example of co-evolution is the growth of a black hole and a pseudobulge, a concentration of starlight near the centre of the galaxy, but in the disk, not extending above it as do stars in a bulge. Pseudobulges may be the accumulation of disk material that migrated towards the galactic centre, some of it tumbling all the way in to feed the black hole's growth. Here, however, we lack a signature. Kormendy *et al.*[1] find that, unlike the case for galaxies with bulges, for pure-disk galaxies with pseudobulges, such as M101, the properties of the disk, pseudobulge and black hole are not closely related. A challenge for the advancing power of theoretical methods[5,6] is to understand this inward migration of matter, and why it preferentially fed the pseudobulge in some galaxies and the black hole in others.

There are roughly equal numbers of nearby large galaxies with and without bulges[7]. For example, the galaxy next to ours, M31, has a prominent bulge and a black-hole mass close to 100 million solar masses, whereas our

REF. 7

NASA/ESA/K. KUNTZ (JHU)/F. BRESOLIN (UNIV. HAWAII)/J. TRAUGER (JPL)/J. MOULD (NOAO)/Y.-H. CHU (UNIV. ILLINOIS, URBANA)/STSCI

Figure 1 | The pure-disk galaxy M101. The spiral arms in this nearby galaxy are in a rotating disk of stars and gas seen nearly face-on. The dark streaks are lanes of dust that absorb starlight. The inset is an enlarged view of the central region, and shows dust lanes extending to the tiny central star cluster. Because dust settles near to the disk, the absorption of starlight shows that most of the stars are close to the disk too. That is, this galaxy does not have the bulge of old stars extending above the disk that is a prominent feature of some galaxies. If there is a black hole in the centre of M101, it is tiny compared with black holes in other galaxies of similar mass. This exemplifies Kormendy and colleagues' argument[1,2] that the dark-matter halo of a galaxy has little influence on the formation of its central black hole. The authors also point out that pure-disk galaxies are not uncommon, and that they managed to grow black holes without possessing the bulge that is thought to funnel the mass that grows the black hole of a galaxy that has a bulge.

Galaxy, which has a similar dark-matter halo, is bulgeless, and the black-hole mass is only a few per cent of that in M31. Therefore, the local environment, which was probably similar for these neighbouring galaxies, does not seem to be a significant factor in determining whether a galaxy acquires a bulge and its attendant black hole.

In theory, galaxies both with and without bulges were growing by the gravitational collection of clumps of matter when the star-formation rate was near its peak. That would suggest that the clumps contained stars; a recent discussion puts roughly comparable masses in stars and gas[6]. So where are these early generations of stars? Not in disks, because there is nothing that would slow the motion of a star to allow it to settle onto a disk. Bulges contain old stars, and it has been suggested that this is where the early stars ended up. But we now see that this is not plausible: why would these old stars have avoided our bulgeless Galaxy and settled instead in the bulge of our neighbour M31? Maybe the old stars are in diffuse stellar haloes. If so, it seems curious that the stellar halo of our Galaxy is much less prominent than that of M31. But more studies of other nearby galaxies will be required to check for inventories of stars that are old enough and abundant enough to account for stars that formed before disks.

Our standard galaxy-formation theory is informed by the hot Big Bang cosmological model. The demanding tests that this model passes[8] show that we must take it seriously, but so far it has not offered much guidance for understanding the evidence Kormendy *et al.*[1,2] present. And because the Universe has surprised us before, I would not ignore the possibility that the cosmological model requires fine adjustment to account for a relatively small detail — the galaxies. ■

P. James E. Peebles *is in the Department of Physics, Princeton University, Princeton, New Jersey 08544, USA.*
e-mail: pjep@princeton.edu

1. Kormendy, J., Bender, R. & Cornell, M. E. *Nature* **469,** 374–376 (2011).
2. Kormendy, J. & Bender, R. *Nature* **469,** 377–380 (2011).
3. Hopkins, A. M. & Beacom, J. F. *Astrophys. J.* **651,** 142–154 (2006).
4. Richards, G. T. *et al. Astron. J.* **131,** 2766–2787 (2006).
5. Brook, C. B. *et al.* Preprint at http://arxiv.org/abs/1010.1004 (2010).
6. Stewart, K. R., Bullock, J. S., Wechsler, R. H. & Maller, A. H. *Astrophys. J.* **702,** 307–317 (2009).
7. Kormendy, J., Drory, N., Bender, R. & Cornell, M. E. *Astrophys. J.* **723,** 54–80 (2010).
8. Komatsu, E. *et al.* Preprint at http://arxiv.org/abs/1001.4538 (2010).

AIDS

Drugs that prevent HIV infection

Two human trials investigate the efficacy of a type of antiretroviral drug — usually used to treat HIV-infected individuals — in preventing HIV infection. The results are heartening.

MARK A. WAINBERG

Each year, several million people become infected with HIV — the agent responsible for AIDS. This formidable statistic highlights the need for drugs that could be given as a preventive measure to vulnerable populations. Such drugs would be as valuable as those currently used to prevent infection after recent possible exposure to the virus and to treat the symptoms of HIV infection. Two papers, one published in *Science*[1] and the other in *The New England Journal of Medicine*[2], now report on a class of antiretroviral drug that can prevent HIV infection in a significant proportion of individuals. Although previous animal studies have shown evidence for such effects[3,4], the current trials provide the first proof of principle for the approach in humans.

The two studies used different modes of drug administration to achieve protection. In the first, a microbicide-based approach carried out in South Africa, Abdool Karim *et al.*[1] asked roughly 450 women who were at risk of acquiring HIV to self-administer a vaginal gel impregnated with the drug tenofovir before sexual intercourse. The authors report a 39% reduction in HIV transmission in these women over 2.5 years compared with a similar number of women who received a placebo gel.

Notably, Abdool Karim and colleagues point to an excellent correlation between adherence to the use of the tenofovir gel and the extent of protection. The levels of protection in women who used the gel at least 80% of the time, 50–80% of the time, or less than 50% of the time were 54%, 38% and 28%, respectively. An added benefit was that a high proportion of the women who used the gel were also protected from infection by an unrelated virus — herpes simplex virus.

Future work will doubtless make use of combinations of antiviral drugs in vaginal gels. For instance, a gel based on both tenofovir and another antiretroviral drug, emtricitabine, should perform better than one based on either agent alone. Indeed, in the second study Grant *et al.*[2] investigate the efficacy of tenofovir and emtricitabine, in this case as a co-formulated, daily administered single pill.

The researchers gave the pill, or a placebo pill, to nearly 2,500 sexually active homosexual men as oral pre-exposure prophylaxis (PREP). Over the following median period of 1.2 years, they documented a 44% reduction in HIV acquisition in the experimental group compared with subjects receiving the placebo. Although Grant and colleagues did not study the efficacy of their pill among vulnerable women, similar results would be expected.

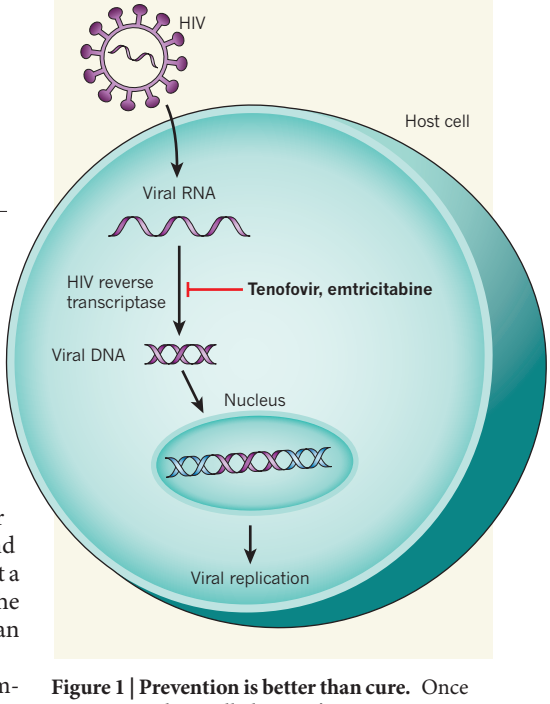


Figure 1 | Prevention is better than cure. Once HIV enters a host cell, the virus's reverse transcriptase enzyme converts its RNA genome into DNA. The viral DNA then becomes integrated into the cellular genome to facilitate HIV replication. Two clinical trials[1,2] show that the drugs tenofovir and emtricitabine — which potently inhibit HIV reverse transcriptase, thereby blocking viral replication — can prevent the initiation of persistent infection in uninfected individuals. These drugs, however, cannot eliminate cells already infected with the virus.

Like Abdool Karim and co-workers, Grant *et al.* note the importance of adherence to the drug regimen: their pharmacological data reveal detectable levels of both tenofovir and emtricitabine in only 3 (9%) of the 34 individuals who became infected, in contrast to 22 (51%) of 43 people who remained uninfected.

This common message of both papers is highly encouraging, and should help to promote adherence among populations at risk.

Another common theme of the two studies reinforces the idea that drugs that quickly reach high levels in the blood and other body compartments are more effective than those that do not: high levels of each of tenofovir and emtricitabine are maintained in the relevant body fluids and tissues for at least 24 hours. The idea has now prompted some investigators to explore the alternative approach of intermittent PREP, whereby tenofovir and emtricitabine are taken only on the morning before the anticipated sexual intercourse, rather than every day. It is noteworthy that, for ethical reasons, all of these studies encourage participants to also use condoms, although generally many of the subjects do not comply.

The antiretroviral drugs used in the current studies[1,2] act on the HIV reverse transcriptase enzyme (Fig. 1). Further research might instead focus on drugs that target other molecules involved in the virus's life cycle; attractive candidates include compounds that block HIV attachment to target cells. One such drug, maraviroc, which is approved for treating HIV infections, has been licensed for use in the development of microbicides, and there is a strong rationale for its use in oral PREP.

It is not surprising that antiviral drugs can prevent sexual transmission of HIV. It has long been known that HIV-infected women who receive such drugs during pregnancy have a greatly reduced risk of transmitting the virus to their babies[5]. Moreover, both oral-PREP and microbicide studies[6,7] in monkeys have shown that either tenofovir or joint administration of tenofovir and emtricitabine can prevent infection with the HIV-related virus SIV. What's more, individuals whose HIV infection is well managed with efficient use of antiretroviral drugs are relatively non-infectious to their sexual partners, even if condoms are not used[8,9].

The new work[1,2], however, could potentially take us beyond the 60% success rate in preventing HIV transmission that has been obtained through male circumcision[10], because the latter procedure remains difficult to implement in low-income countries. The use of a pill may be the easiest option of all.

Some dangers remain, nonetheless. For instance, if antiviral drugs — administered as either microbicides or oral PREP — are given to people who are already infected with HIV but not yet diagnosed, it might result in drug resistance. This is because, given alone, neither tenofovir nor co-administered tenofovir and emtricitabine can fully suppress viral replication, but can potentially create an opportunity for selection and subsequent transmission of drug-resistant HIV strains[11]. This issue may be more relevant for oral PREP than for microbicides, particularly if the gel formulations used for the latter can prevent the systemic absorption of the drugs that could lead to the selection of resistant viruses. This potential problem notwithstanding, the results of these two trials represent marked progress in research to prevent HIV infection. ■

Mark A. Wainberg *is at the McGill University AIDS Centre, Lady Davis Institute, Jewish General Hospital, Montreal, Quebec H3T 1E2, Canada.*
e-mail: mark.wainberg@mcgill.ca

1. Abdool Karim, Q. *et al. Science* **329,** 1168–1174 (2010).
2. Grant, R. M. *et al. N. Engl. J. Med.* **363,** 2587–2599 (2010).
3. Van Rompay, K. K. A. *et al. J. Infect. Dis.* **184,** 429–438 (2001).
4. Youle, M. & Wainberg, M. A. *AIDS* **17,** 937–938 (2003).
5. De Cock, K. M. *et al. J. Am. Med. Assoc.* **283,** 1175–1182 (2000).
6. García-Lerma, J. G. *et al. PLoS Med.* **5,** e28 (2008).
7. García-Lerma, J. G. *et al. Sci. Transl. Med.* **2,** 14ra4 (2010).
8. Quinn, T. C. *et al. N. Engl. J. Med.* **342,** 921–929 (2000).
9. Vernazza, P. L. *et al. AIDS* **11,** 1249–1254 (1997).
10. Auvert, B. *et al. PLoS Med.* **2,** e298 (2005).
11. Wainberg, M. A. & Brenner, B. G. *Viruses* **2,** 2493–2508 (2010).

IMAGING

Spot the hotspot

Plasmonic hotspots — nanometre-sized crevices that permit the detection of single molecules — are too small to be imaged with conventional microscopes. They can now be probed using super-resolution fluorescence microscopy. SEE LETTER P.385

MARTIN MOSKOVITS

When light is shone on molecules placed on appropriately nanostructured gold or silver surfaces, the light scattered from the molecules is often amplified by factors of a million — or even a billion. This amplification effect is the working principle of surface-enhanced Raman spectroscopy (SERS)[1,2], and allows the technique to detect and identify very small samples of molecules, often down to single molecules. Among researchers, there is a consensus that the effect arises because the nanostructured metals can harvest the incoming light in their environments — much like an antenna collects radio waves — and concentrate the light's electromagnetic energy in 'plasmonic hotspots', often nanoscale clefts, gaps and fissures. However, although the optical field within a hotspot has been the subject of numerous calculations and simulations, direct measurements of the field's profile have been hard to carry out. On page 385 of this issue, Cang *et al.*[3] show how they have managed to do just that.

The most effective hotspots — those yielding the highest light enhancement — are a few nanometres in diameter. However, the normal image resolution of a conventional optical microscope, which is limited to half the wavelength of the incident light, is tenfold too coarse to probe such small areas. Even unconventional imaging devices, such as near-field scanning optical microscopes, are limited by the size of their probes, which rarely fall below 30 nanometres[4].

Enter Cang and colleagues[3], who adapted the technique of single-molecule, super-resolution optical fluorescence microscopy[5] to probe SERS hotspots. This method makes use of the fact that, although the diameter of the apparent image of any given molecule is much larger than the molecule (in fact, it is at least as large as half the wavelength of the light emitted by the molecule), if the individual molecules of a molecular ensemble can be made to fluoresce one at a time, the diameter of the aggregate image becomes much smaller. Indeed, the technique can yield image resolution of less than 2 nm.

To image single hotspots on the surface of a metal, Cang *et al.* submerged the metal in a solution of fluorescent dye molecules, and by choosing the appropriate concentration they ensured that, on average, molecules would arrive at the hotspot one at a time. On binding to the hotspot, a molecule's fluorescence increases greatly and appears as a bright spot, the intensity of which is a measure of the enhancement. The adsorbed dye molecule's fluorescence dies out within hundreds of milliseconds of its arrival at the hotspot, and the hotspot is then ready for the next adsorption event.

In this way, Cang and colleagues[3] were able to image the fluorescence enhancement profile of single hotspots as small as 15 nm with an accuracy of 1–2 nm. A recent study by Stranahan and Willets[6], which was based on an imaging technique similar to that used here, has also reported single-hotspot images, but with a resolution of the order of 10 nm.

For hotspots in silver nanostructures, Cang *et al.* measured fluorescence-intensity enhancements in excess of 130. Fluorescence enhancements are significantly lower than those of SERS for two reasons. First, in fluorescence only the incident field is enhanced,

whereas SERS enhancement derives from the amplification of both the incident and scattered fields — in fact, the SERS enhancement effect is really enhanced scattering of the molecules' Raman emission by the metal. Second, the fluorescence of molecules in close proximity to a metal is quenched to some extent[7].

Cang *et al.* also demonstrated that the SERS enhancement effect correlates inversely with the size of the hotspot. That is, the optical field was most strongly concentrated in the smallest hotspots, an effect that had previously been predicted by simulations but not measured directly. Taken together with the earlier report of Stranahan and Willets[6], the authors' results[3] provide direct evidence for optical-field concentration in hotspots, which has previously been supported by only indirect — although compelling — evidence. A current challenge for SERS is the engineering of substrates with a very large density of hotspots, which could function as highly sensitive biomolecular sensors. With the demonstrated ability to image hotspots directly, that goal seems within closer reach. ■

Martin Moskovits *is in the Department of Chemistry and Biochemistry, University of California, Santa Barbara, California 93106–9510, USA.*
e-mail: moskovits@chem.ucsb.edu

1. Jeanmaire, D. L. & Van Duyne, R. P. *J. Electroanal. Chem.* **84,** 1–20 (1977).
2. Moskovits, M. *Rev. Mod. Phys.* **57,** 783–826 (1985).
3. Cang, H. *et al. Nature* **469,** 385–388 (2011).
4. Betzig, E. & Trautman, J. K. *Science* **257,** 189–195 (1992).
5. Hell, S. W. in *Single Molecule Spectroscopy in Chemistry, Physics and Biology* (eds Gräslund, A., Rigler, R. & Widengren, J.) 365–398 (Springer, 2010).
6. Stranahan, S. M. & Willets, K. A. *Nano Lett.* **10,** 3777–3784 (2010).
7. Chance, R. R., Prock, A. & Silbey, R. *Adv. Chem. Phys.* **37,** 1–65 (1978).

EVOLUTIONARY BIOLOGY

Farming writ small

Social slime moulds graze on bacteria, but save some for transmission in their spores. Strains practising this primitive form of farming coexist with non-farmer strains in an intriguing cost–benefit equilibrium. SEE LETTER P.393

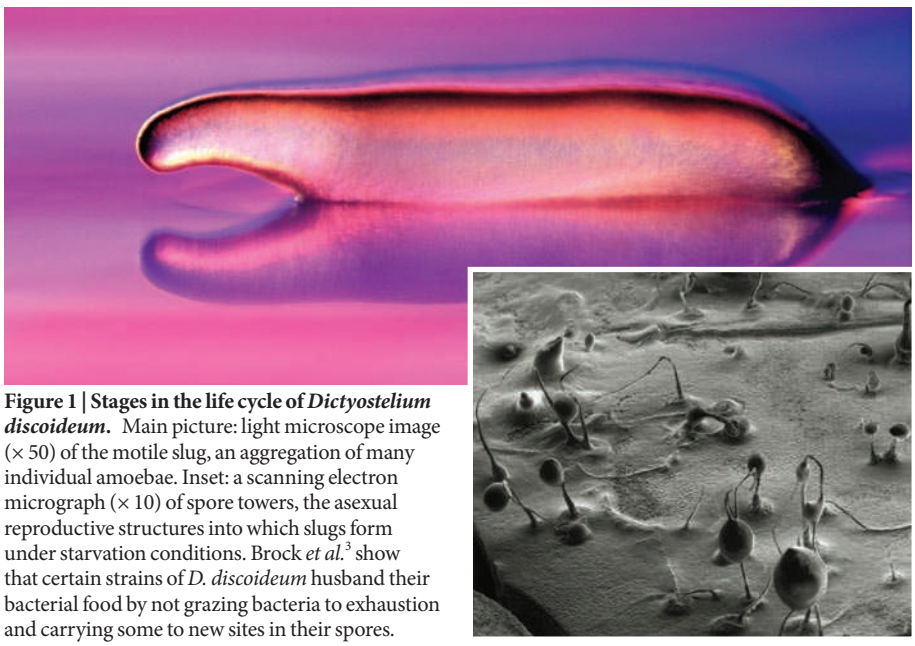
JACOBUS J. BOOMSMA

In the space of just a few millennia, the domestication of crops and livestock enabled bands of human hunter–gatherers to transform into the computerized, space-exploring societies of today. A predictable food supply ultimately turned out to be a great benefit, but it has remained unclear how the initial transitions to farming prevailed in the face of the inevitable infrastructural costs and vulnerability to disease[1]. Societies of fungus-growing ants and termites, and wood-boring ambrosia beetles, have also evolved advanced agriculture[2]; here, too, the absence of surviving lineages in which farming and non-farming co-occur make it difficult to evaluate how these transitions became irreversible. Brock *et al.* (page 393 of this issue)[3] now add to the picture. Their demonstration that *Dictyostelium* slime moulds practise bacterial husbandry sheds light on the trade-offs governing incipient domestication.

Dictyostelium exist as soil-dwelling, unicellular amoebae that can reproduce sexually[4]. But they are best known for their social reproduction, during which cells aggregate and form a motile multicellular slug that subsequently produces a fruiting body containing asexual spores[5] (Fig. 1). This happens whenever the amoebae have exhausted their local bacterial prey, on which they feed in a quite catholic manner[3]. One species — *Dictyostelium discoideum* — is a particularly important model system, both for developmental biologists studying the origin of multicellularity[5] and for evolutionary biologists interested in the conflicts and cooperation that occur under the conditions of kin selection[6]. Brock and colleagues[3] show that a significant fraction of *D. discoideum* spores carry bacteria to inoculate new habitat with food, but that such husbandry has remained clone-specific, with only about one-third of the strains examined practising it.

The authors' molecular and experimental evidence indicates that strains are either farmers or non-farmers and that spore-borne bacteria confer substantial fitness benefits on *Dictyostelium* when spores land in patches without suitable food. However, this form of farming has significant costs because patches cannot be exhaustively grazed before cells aggregate for reproduction. Slugs loaded with bacteria also travel less far, and investments in co-transmission do not pay off when spores germinate in places where food bacteria are present. These context-dependent benefits make the *Dictyostelium* farming polymorphism an example of a mixed evolutionarily stable strategy[7]. This is a fascinating finding, because the *Dictyostelium*–bacterial symbiosis is evidently driven by mutualistic advantages, despite the obvious risks of bacterial exploitation of the dispersal opportunities provided by the hosts.

Dictyostelium slime moulds belong to an ancient group, sister to fungi and animals combined[8]. So, despite the long evolutionary timescale, why have the benefits of farming remained such a mixed bag? It seems that what has stalled developments towards unambiguously profitable farming is that slime moulds did not become specialized on a single bacterial species, which could then have co-evolved and co-speciated with its host and lost its free-living stages. Slime mould lineages did adapt to farming, but their bacterial 'crops' have remained under selection for independent growth. There is no substrate facilitation, growth reinforcement, or removal of competitors or weeds, as



CAROLINA BIOLOGICAL/VISUALS UNLIMITED/CORBIS

D. SCHARF/SPL

Figure 1 | Stages in the life cycle of *Dictyostelium discoideum*. Main picture: light microscope image (× 50) of the motile slug, an aggregation of many individual amoebae. Inset: a scanning electron micrograph (× 10) of spore towers, the asexual reproductive structures into which slugs form under starvation conditions. Brock *et al.*[3] show that certain strains of *D. discoideum* husband their bacterial food by not grazing bacteria to exhaustion and carrying some to new sites in their spores.

seen in the classic fungus-farming symbioses[2] and in the more recently discovered incipient practices of fungus farming by *Littoraria* snails[9] and red-alga farming by *Stegastes* damselfish[10].

The damselfish mutualism has led to at least one case of monoculture in which the crops no longer have free-living relatives. The fact that slime mould husbandry has not achieved this status underlines the point that parent–offspring transmission is insufficient to install absolute co-dependency in a mutualism[11]. Some form of monoculture farming is apparently essential to make symbionts put all their eggs in a host's basket, because being eaten is profitable only when it benefits clone mates that are nursed and dispersed. Monocultural commitment makes these kin-selected benefits consistent, paving the way for mutual coadaptation, irrespective of symbionts being acquired from the environment rather than being inherited[12]. The limitations of bacterial husbandry in slime moulds[3] therefore clarify a major cornerstone of our general understanding of mutualistic interactions. They also invite further study to unravel the molecular mechanisms that allow or prevent bacterial transmission, and to establish the dynamics of food transmission in slugs that are genetic mixtures of several strains[6].

The *Dictyostelium* symbiosis presents an interesting analogy to culturally adjustable human subsistence farming in its various contemporary and historical combinations with hunter–gatherer strategies[1]. In both slime moulds and humans, farmers did not become reproductively isolated from non-farmers, nor did crops or livestock lose the possibility of hybridizing with wild relatives, as has happened in the specialized insect fungus-farming symbioses[2,12]. The slime moulds may have insufficient multicellular complexity to evolve specialized nurturing traits for particular crops, whereas our own species lacked evolutionary time and consistent selection for extreme crop specialization. Neither of these constraints applied to the farming societies of ants and termites.

Although *Dictyostelium* do not actively rear their crops, they may well possess unknown adaptations that, if revealed, would illuminate fundamental questions of conflict and cooperation across species boundaries. The ancestors of these slime moulds were among the earliest colonizers of terrestrial habitats, so the history of this bacterial husbandry symbiosis may go back further than any other farming system. Mapping farming practices on a large-scale evolutionary tree of the slime moulds would therefore be a worthy objective. If this farming symbiosis turns out to be ancient, our new understanding of *Dictyostelium* biology could be summed up in a lyric of the rock band Metallica. To paraphrase: wherever they roamed, they redefined the unknown, by themselves but not alone. ■

Jacobus J. (Koos) Boomsma *is in the Centre for Social Evolution, Department of Biology, University of Copenhagen, 2100 Copenhagen, Denmark.*
e-mail: jjboomsma@bio.ku.dk

1. Diamond, J. *Nature* **418,** 700–707 (2002).
2. Mueller, U. G. *et al. Annu. Rev. Ecol. Evol. Syst.* **36,** 563–595 (2005).
3. Brock, D. A., Douglas, T. E., Queller, D. C. & Strassmann, J. E. *Nature* **469,** 393–396 (2011).
4. Bloomfield, G., Skelton, J., Ivens, A., Tanaka, Y. & Kay, R. R. *Science* **330,** 1533–1536 (2010).
5. Kessin, R. H. *Dictyostelium: Evolution, Cell Biology, and the Development of Multicellularity* (Cambridge Univ. Press, 2001).
6. Gilbert, O. M. *et al. Proc. Natl Acad. Sci. USA* **104,** 8913–8917 (2007).
7. Maynard Smith, J. *Proc. R. Soc. Lond. B* **205,** 475–488 (1979).
8. Baldauf, S. L. *et al. Science* **290,** 972–977 (2000).
9. Silliman, B. R. & Newell, S. Y. *Proc. Natl Acad. Sci. USA* **100,** 15643–15648 (2003).
10. Hata, H., Watanabe, K. & Kato, M. *BMC Evol. Biol.* **10,** 185 (2010).
11. Herre, E. A., Knowlton, N., Mueller, U. G. & Rehner, S. A. *Trends Ecol. Evol.* **14,** 49–53 (1999).
12. Aanen, D. K. *et al. Science* **326,** 1103–1106 (2009).

IMMUNOLOGY

Peptide gets in shape for self-defence

The transformation of tadpole to frog and of caterpillar to butterfly are two of the more obvious examples of metamorphosis. But molecular shape-shifting may occur in each of us as part of our innate antibacterial defence system. SEE LETTER P.419

ROBERT I. LEHRER

Among the immune mediators that fight microorganisms within us, one is human β-defensin 1 (hBD-1). This peptide, which was first described[1] in 1995, is continually expressed in skin and epithelial cells throughout the body[2]. But because its direct antibacterial properties are modest, the reason it stands guard at interfaces between microbe-laden environments — such as the colon or skin — and their adjacent, normally sterile tissues has remained enigmatic. On page 419 of this issue, Schroeder *et al.*[3] show that the mild antibacterial activity of hBD-1 changes drastically after it undergoes a chemically induced change in shape.

The peptide backbone of hBD-1 is folded into a well-defined structure that is held together by three internal disulphide bonds[4]. Schroeder *et al.* find that an enzyme called thioredoxin reductase can sever these disulphide bonds in a reduction reaction. The reduced hBD-1 molecule undergoes a profound change in shape (Fig. 1, overleaf) that allows it to kill some Gram-positive bacteria, against which its normally oxidized form is powerless. The authors' electron micrographs of the slain bacteria show changes that might lead a forensically inclined microbiologist to wonder whether reduced hBD-1 induced the bacteria to self-destruct, by triggering their latent autolysis (self-breakdown) systems[5].

Schroeder and co-workers' observations[3] are unlikely to be merely a 'test-tube' phenomenon. They demonstrate that in human epithelia, oxidized hBD-1 and thioredoxin reductase colocalize with reduced hBD-1.

If shape change alone imparted the expanded antimicrobial range of hBD-1, then analogues of this peptide in which cysteine residues are replaced by other amino acids should also show enhanced function, because the cysteine-free peptides would be unable to form shape-restraining disulphide bonds. But Schroeder *et al.* report that such analogues are ineffective antibiotics. Evidently, there is something special about the cysteine residues of hBD-1, but exactly what remains unknown.

There is also something special about having a positive charge. Full-length hBD-1 has 36 amino-acid residues, including six cysteines, one negatively charged aspartic acid and five positively charged residues (four lysines and one arginine). If the partial charge of its single histidine residue is ignored, hBD-1 has a net charge of +4. This charge is concentrated in its carboxy-terminal octapeptide — arginine-glycine-lysine-alanine-lysine-cysteine-cysteine-lysine. A truncated variant of hBD-1 lacking the last seven of these residues shows no activity, but a seven-residue peptide lacking only the initial arginine of the octapeptide does.

It is ironic that the bacteria that Schroeder *et al.* find to be susceptible to reduced hBD-1 belong to either the lactobacilli or bifidobacteria genera — organisms that are generally considered[6] to be health-promoting probiotics rather than potential pathogens. This observation, however, should be considered a proof of concept, rather than a serious colonic conundrum.

At least 1,000 species of bacteria reside in the colon of a healthy adult, and a single gram of faeces may contain up to 10^{12} bacteria. Unlike bifidobacteria, most bacterial species residing

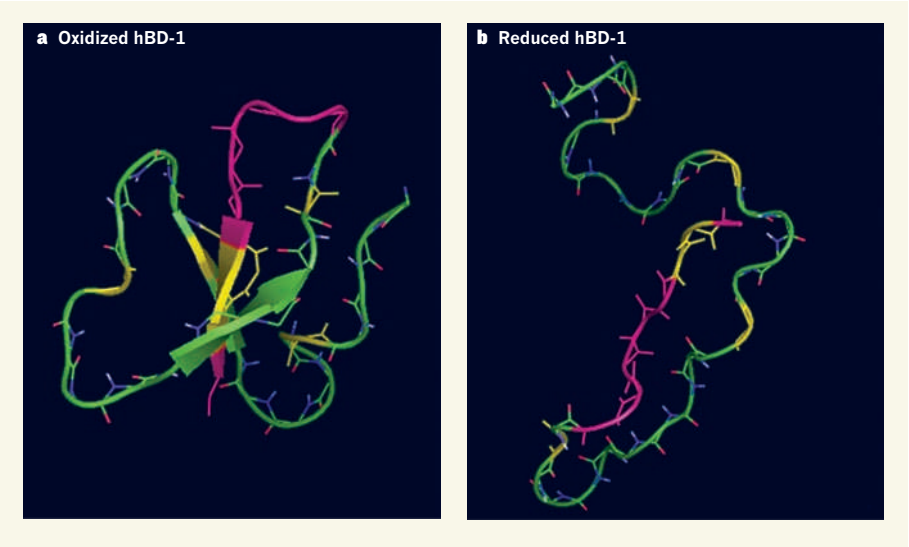


Figure 1 | Computer-generated structures of human β-defensin 1. **a**, In its oxidized form, the peptide human β-defensin 1 (hBD-1) contains three disulphide bonds. The peptide backbone is shown in green except for cysteine residues (yellow) and the six non-cysteine carboxy-terminal residues (pink). The broad arrows represent β-sheet components found only in the oxidized form. **b**, The reduced hBD-1 structure was generated from the lowest-energy conformation of oxidized hBD-1, after breaking its disulphide bonds and causing it to assume a random conformation. (Structure generated by Alan J. Waring, Univ. California, Los Angeles.)

in the colon cannot be grown in culture, and their presence can only be disclosed using various 'gene-sniffing' techniques. It could be, therefore, that the effects of reduced hBD-1 on probiotic bacteria are simply collateral damage on these harmless bystanders by a defence system that also targets less-well-intentioned intestinal residents or transients. Alternatively, even probiotics may require surveillance to keep them from overstepping their boundaries.

Ideally, the activity of an antibiotic should be examined in a defined medium, the composition of which closely resembles, or precisely replicates, the *in vivo* environment. With precise simulation of the colonic content being too challenging to contemplate, it would be informative to learn how defined factors — such as pH and, in particular, salinity — affect the activity of reduced hBD-1 *in vitro*. Another interesting experiment would be to test the antibacterial activity of mixtures of reduced and oxidized hBD-1, because clearly such mixtures occur *in vivo*.

It remains unknown whether Schroeder and colleagues' results are unique to hBD-1 or whether they are also true for other defensin peptides. Defensins and defensin-like peptides are fairly universal participants in host defence against infection[7]: they occur in plants, fungi, invertebrates and vertebrates. Vertebrates have three subfamilies of defensins (designated α, β and θ)[8], the members of which consist exclusively of cationic peptides with six cysteines and three disulphide bonds, which also provide resistance to premature proteolytic digestion.

Although more than 20 genes have been identified[9] that encode hBDs, only hBDs 1–4 have received extensive attention. The net positive charge of these four peptides varies from +4 for hBD-1 to an astounding +11 for hBD-3, whose eight carboxy-terminal residues alone carry a net charge of +6. From previous work[10] on hBD-3, its high net positive charge contributes substantially to the peptide's ability to kill bacteria or fungi such as *Candida albicans*. This is especially true when the assays are performed in media of low ionic strength.

Given the extreme cationicity and high intrinsic activity of oxidized hBD-3, it is not surprising that when Schroeder *et al.*[3] removed its disulphide bonds, they did not detect improved activity of this peptide against bifidobacteria. After all, a bacterium can only be killed once. For thioredoxin reductase to empower hBD-3 to do so twice would be a *reductio ad absurdum*. ■

Robert I. Lehrer *is in the Department of Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California 90045, USA.*
e-mail: rlehrer@mednet.ucla.edu

1. Bensch, K. W., Raida, M., Mägert, H.-J., Schulz-Knappe, P. & Forssmann, W.-G. *FEBS Lett.* **368,** 331–335 (1995).
2. Zhao, C., Wang, I. & Lehrer, R. I. *FEBS Lett.* **396,** 319–322 (1996).
3. Schroeder, B. O. *et al. Nature* **469,** 419–423 (2011).
4. Schibli, D. J. *et al. J. Biol. Chem.* **277,** 8279–8289 (2002).
5. Sahl, H.-G. *et al. J. Leukoc. Biol.* **77,** 466–475 (2005).
6. Kleerebezem, M. & Vaughan, E. E. *Annu. Rev. Microbiol.* **63,** 269–290 (2009).
7. Wong, J. H., Xia, L. & Ng, T. B. *Curr. Protein Pept. Sci.* **8,** 446–459 (2007).
8. Selsted, M. E. & Ouellette, A. J. *Nature Immunol.* **6,** 551–557 (2005).
9. Schutte, B. C. *et al. Proc. Natl Acad. Sci. USA* **99,** 2129–2133 (2002).
10. Hoover, D. M. *et al. Antimicrob. Agents Chemother.* **47,** 2804–2809 (2003).

CHEMICAL BIOLOGY

Catalytic detoxification

Protein engineering of an enzyme that catalytically detoxifies organophosphate compounds in the body opens up fresh opportunities in the search for therapeutic protection against nerve agents used in chemical warfare.

FRANK M. RAUSHEL

Organophosphates are among the most toxic compounds that have been chemically synthesized. Since the discovery of their biological activity in the 1930s, these compounds have found use as broad-spectrum insecticides for agricultural and domestic applications. But organophosphates have also been developed as chemical-warfare agents, including VX and the 'G-agents' (such as sarin, soman and cyclosarin). Because these compounds are relatively easy to synthesize, their use by international terrorist groups is a serious threat. Current protocols for the prevention and treatment of organophosphate poisoning are largely ineffective, and so new strategies are desperately needed. Reporting in *Nature Chemical Biology*, Gupta *et al.*[1] describe an approach that might one day find use in preventing organophosphate poisoning.

Organophosphates are highly toxic because they rapidly inactivate acetylcholinesterase (AChE), an enzyme required for nerve function (Fig. 1). AChE breaks down (hydrolyses) acetylcholine, a neurotransmitter that relays nerve impulses to muscles and other organs. Organophosphates form a covalent bond to a serine amino-acid residue in the active site of AChE, stopping the enzyme from functioning. The subsequent build-up of acetylcholine blocks cholinergic nerve impulses, leading to paralysis, suffocation and death.

Various prophylactic approaches have been developed to diminish the toxic effect of organophosphates. Atropine, for example, is a competitive antagonist for muscarinic acetylcholine receptors — it blocks the action

of acetylcholine, thereby reducing the effective concentration of the neurotransmitter. Alternatively, chemicals such as pralidoxime react with AChE–organophosphate adducts to regenerate catalytically active AChE.

A relatively new approach for reducing the concentration of organophosphates in the blood is to inject human butyrylcholinesterase (BChE) directly into the bloodstream of a poisoned individual[2]. This enzyme reacts with organophosphates in the same way as AChE, and thus acts as a selective scavenger for the nerve agents. But a problem with this approach is that the scavenging reaction is stoichiometric — one BChE molecule is required to scavenge one molecule of organophosphate. This means that a substantial amount of the enzyme must be injected into the body to reduce a lethal dose to non-toxic levels: approximately 350 milligrams of BChE are required to detoxify every milligram of cyclosarin, for example, because the molecular mass of the enzyme is much greater than that of the nerve agent.

Gupta *et al.*[1] now report genetically modified enzymes that hydrolyse organophosphates. Notably, the nerve agents bind to these enzymes as substrates (non-covalently and reversibly), rather than as potent inactivators (which bind covalently and irreversibly). The enzymes therefore behave as catalysts for organophosphate clearance — each enzyme molecule destroys thousands of molecules of a nerve agent, thus reducing the amount of enzyme required to detoxify a lethal dose.

The authors' work builds on an earlier study[3] in which the active site of a bacterial phosphotriesterase enzyme was optimized by protein engineering to effectively catalyse the hydrolysis of a wide range of organophosphates, including sarin, soman and cyclosarin. Expression of this enzyme in caterpillars reduced the lethal effects of paraoxon[4], an insecticide whose active form is an organophosphate. It is unlikely, however, that a protein of bacterial origin could be used as an effective therapeutic agent for organophosphate toxicity in humans.

Gupta *et al.*[1] have made great strides towards solving this problem by using an enzyme largely of human origin. They worked with a serum paraoxonase enzyme, PON1, which catalyses the hydrolysis of lactones (cyclic molecules often found in nature), but which also catalyses the slow hydrolysis of a range of organophosphate nerve agents. Unfortunately, PON1 isn't stable enough to be expressed and manipulated in bacterial cells, as is required for engineering its catalytic properties. The authors partly overcame this challenge by using a hybrid enzyme[5] that combined parts of human and rabbit PON1 enzymes.

If an enzyme is to be an effective catalyst for the hydrolysis of organophosphate nerve agents, the rate constant (k_{cat}/K_m, a measure of the speed of a chemical reaction) for the process must exceed 10^7 M^{-1} min^{-1}. Another consideration to take into account is that the G-agents are chiral molecules — they form isomers known as enantiomers that are mirror images of each other. Only one of the two enantiomers (the S_P enantiomer) is toxic. Frustratingly, Gupta and colleagues found that their hybrid PON1 enzyme primarily hydrolysed the non-toxic isomer (the R_P enantiomer) of an analogue of cyclosarin, and that k_{cat}/K_m for the hydrolysis of the S_P enantiomer was less than 200 M^{-1} min^{-1}.

The authors therefore subjected the hybrid PON1 to a series of directed-evolution and rational-design experiments, in which alterations were made to the amino-acid sequence of the protein and the resulting mutants were screened to assess their effectiveness in hydrolysing organophosphates. To identify the most effective mutants, the researchers developed a sorting procedure that compartmentalized individual bacteria expressing the mutant enzymes in emulsion droplets. By adding a substrate to the droplets that produces a fluorescent compound when hydrolysed by the enzymes, those bacteria expressing active mutants were easily detected and isolated. The most active mutants were then subjected to further rounds of alterations and screening. After several generations of enzymes had been produced, the authors identified a mutant for which k_{cat}/K_m for the hydrolysis of the toxic isomer of a cyclosarin analogue exceeded 10^7 M^{-1} min^{-1}. This represents an enhancement of 100,000 over the activity of the starting enzyme.

The authors went on to show that the newly evolved enzyme acts as a prophylactic against G-agent exposure in mice. When they injected the animals with the recombinant enzyme at a dose of 2.2 milligrams per kilogram of animal body weight one hour before giving them a lethal dose of a cyclosarin analogue, 75% of the mice were still alive after 24 hours. By contrast, none of the control mice survived. However, when mice were challenged with a lethal dose of the organophosphate six hours after injection of the recombinant enzyme, the survival rate dropped to 50%. No protection was afforded when the mice were challenged after 24 hours.

Gupta and colleagues' study nicely demonstrates that a predominantly human enzyme can be engineered to provide significant therapeutic protection against lethal exposure to nerve agents. The catalytic detoxification of organophosphate nerve agents is clearly superior to the stoichiometric detoxification currently afforded by human BChE, because much lower doses of the modified PON1 enzyme are needed for a therapeutic effect. But for the authors' approach to be truly practical, the duration of the catalytic scavenger's activity in humans must be increased beyond the few hours that the current experiments indicate. In addition, the substrate specificity and catalytic efficiencies of scavengers must be expanded and enhanced for activity against other G-agents and against the even more lethal VX. One might also envisage the development of prophylactics to protect people from exposure to agricultural organophosphate pesticides, which have caused far more medical problems over time than the use of military nerve agents. ■

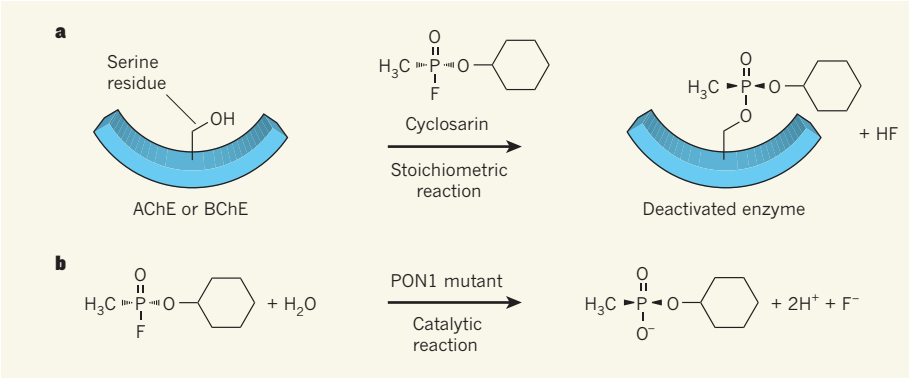


Figure 1 | Enzyme reactions of organophosphate nerve agents. **a**, *In vivo*, organophosphate nerve agents such as cyclosarin react to form a covalent bond with a serine amino-acid residue in the active site of acetylcholinesterase (AChE). This inactivates AChE, preventing it from hydrolysing the neurotransmitter acetylcholine, leading to suffocation and ultimately death. Butyrylcholinesterase (BChE) reacts with organophosphates in the same way, and can be injected into the bloodstream to scavenge the nerve agents, reducing their concentration to non-toxic levels. Because BChE is a stoichiometric scavenger, large concentrations are required for it to be therapeutically effective. **b**, Gupta *et al.*[1] have engineered serum paraoxonase (PON1) to catalyse the rapid hydrolysis of organophosphates. One PON1 molecule catalyses thousands of hydrolysis reactions, and so the amount of enzyme required to reduce the concentration of organophosphates to non-toxic levels is much lower than that needed for BChE.

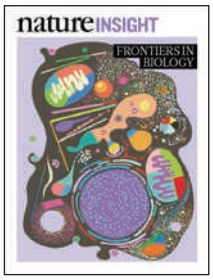
Frank M. Raushel *is in the Department of Chemistry, Texas A&M University, College Station, Texas 77842, USA.*
e-mail: raushel@tamu.edu

1. Gupta, R. D. *et al. Nature Chem. Biol.* doi:10.1038/nchembio.510 (2011).
2. Ashani, Y. & Pistinner, S. *Toxicol. Sci.* **77,** 358–367 (2004).
3. Tsai, P.-C. *et al. Biochemistry* **49,** 7979–7987 (2010).
4. Dumas, D. P. & Raushel, F. M. *Experientia* **46,** 729–731 (1990).
5. Aharoni, A. *et al. Proc. Natl Acad. Sci. USA* **101,** 482–487 (2004).

nature INSIGHT

FRONTIERS IN BIOLOGY

20 January 2011 / Vol 469 / Issue No 7330



Cover illustration by Nik Spencer

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel: +44 (0) 20 7833 4000
e: nature@nature.com

nature publishing group

The *Nature* Insight 'Frontiers in Biology' aims to cover timely and important developments in the broader field of biology, ranging from the subcellular to the organismal level, and including molecular mechanisms and biomedicine. The reviews in this Insight discuss in turn specific aspects of the aetiology of cancer, the response to infection and inflammatory disease, cardiovascular development and disease, and epigenetic mechanisms.

In the first review, Jane Visvader considers our current knowledge of the nature of the cells that give rise to cancer — the cells of origin. Tumours are heterogeneous in that different tumours within one organ can have different phenotypes, and individual tumours are made up of phenotypic and functionally heterogeneous cancer cells. The cellular origins of cancer are the subject of considerable debate as their influence can explain at least part of this heterogeneity. The article discusses key questions and concepts, as well as the implications for early cancer diagnosis and prevention.

The review by Beth Levine, Noboru Mizushima and Herbert Virgin covers new developments at the interface between autophagy and immunology. Autophagy is a protective cellular process in which proteins, organelles and pathogens are sequestered in a double-membrane structure before delivery to the lysosome for degradation. Defects in the function of the autophagy pathway, and also impaired autophagy function that is independent of degradation, have been linked to the pathogenesis of infectious diseases and inflammatory syndromes.

Eric Small and Eric Olson highlight recent progress in our understanding of the role of microRNAs in regulating cardiovascular function and disease mechanisms, and discuss the potential opportunities for new microRNA-based diagnostics and therapeutics.

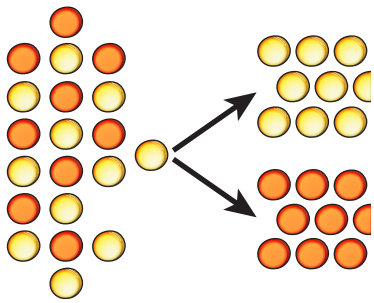
The final review by Raphaël Margueron and Danny Reinberg provides an update on our knowledge about the histone methyltransferase Polycomb repressive complex 2 (PRC2) in regulating chromatin and gene expression, and its effects on development and cancer.

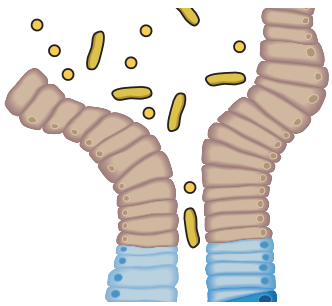
We hope that you will find the articles informative and stimulating.

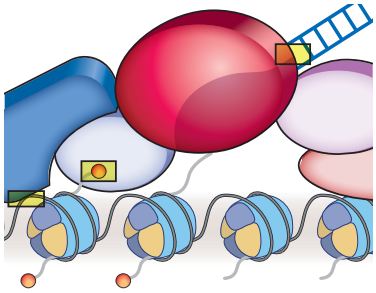
Alex Eccleston, Barbara Marte, Deepa Nath & Clare Thomas
Senior Editors

CONTENTS

REVIEWS







REVIEW

doi:10.1038/nature09781

Cells of origin in cancer

Jane E. Visvader[1,2]

Both solid tumours and leukaemias show considerable histological and functional heterogeneity. It is widely accepted that genetic lesions have a major role in determining tumour phenotype, but evidence is also accumulating that cancers of distinct subtypes within an organ may derive from different 'cells of origin'. These cells acquire the first genetic hit or hits that culminate in the initiation of cancer. The identification of these crucial target cell populations may allow earlier detection of malignancies and better prediction of tumour behaviour, and ultimately may lead to preventive therapies for individuals at high risk of developing cancer.

Remarkably, the oldest description and surgical treatment of cancer dates back to 1600 BC in Egypt, where the papyrus described eight cases of tumours occurring on the breast and their treatment by cauterization. Today, cancer is a leading cause of death worldwide, with the number of deaths from cancer projected to increase markedly, owing partly to an ageing global population. This immense cancer burden demands strategies that permit earlier detection, better stratification of tumours to guide therapy and the development of effective preventive therapies. Targeted therapeutic strategies that suppress the progression of preneoplastic cells towards the malignant state hold great promise for circumventing the huge challenges associated with the treatment of late-stage disease.

Tumours show marked heterogeneity in their cellular morphology, proliferative index, genetic lesions and therapeutic response. The molecular and cellular mechanisms underpinning tumour heterogeneity remain central questions in the cancer biology field. Key issues include whether the different subtypes of cancer reflect a distinct 'cell of origin', the extent to which the genetic mutational profile contributes to tumour phenotype and the nature of the relationship between the cell of origin and the cancer stem cell. This Review focuses on the strategies used to identify cells of origin, the impact of these cells on cancer cell fate and behaviour, and the implications for the development of improved prognostic markers and preventive therapies.

Cell-of-origin and cancer stem-cell concepts are distinct

It is important to note that the cell of origin, the normal cell that acquires the first cancer-promoting mutation(s), is not necessarily related to the cancer stem cell (CSC), the cellular subset within the tumour that uniquely sustains malignant growth. That is, the cell-of-origin and CSC concepts refer to cancer-initiating cells and cancer-propagating cells, respectively (Fig. 1). Although the tumour-initiating cell and the CSC have been used interchangeably, the tumour-initiating cell more aptly denotes the cell of origin. There is considerable evidence that several diverse cancers, both leukaemias and solid tumours, are hierarchically organized and sustained by a subpopulation of self-renewing cells that can generate the full repertoire of tumour cells (both tumorigenic and non-tumorigenic cells)[1]. The cell of origin, the nature of the mutations acquired, and/or the differentiation potential of the cancer cells are likely to determine whether a cancer follows a CSC model. In most instances, the phenotype of the cell of origin may differ substantially from that of the CSC.

Tumour heterogeneity

Phenotypic and functional heterogeneity are hallmarks of cancers arising in several organs. Variability can occur between tumours arising in the same organ (intertumoral heterogeneity), leading to the classification of discrete tumour subtypes. These subtypes are typically characterized by their molecular profile, together with their morphology and expression of specific markers (such as hormone and growth-factor receptors). Variation also occurs within individual tumours (intratumoral heterogeneity), in which the tumour cells often have a range of functional properties and a diverse expression of markers. For example, the proportion of cells that express the oestrogen receptor within a patient's breast tumour can vary extensively, from 1% to 100%. The CSC and clonal-evolution models have been put forward to account for intratumoral heterogeneity and intrinsic differences in tumour-regenerating capacity (reviewed in refs 1 and 2). Interestingly, despite the heterogeneous nature of tumours, the histopathology and gene-expression profiles of tumours arising in patients often remain relatively stable during progression from localized disease to metastatic and even end-stage disease[3,4].

Two main mechanisms have been conceptualized to explain intertumoral heterogeneity: different genetic or epigenetic mutations occurring within the same target cell result in different tumour phenotypes (Fig. 2a), and different tumour subtypes arise from distinct cells within the tissue that serve as the cell of origin (Fig. 2b). It is important to note that these cellular and molecular mechanisms are not mutually exclusive, but can act together to determine tumour histopathology and behaviour. In addition, extrinsic mechanisms may be involved in generating tumour heterogeneity, because interactions between tumour cells and the stromal micro-environment are a crucial determinant of malignant growth[5]. Several studies on human cancers and mouse models have highlighted the importance of specific genetic aberrations in contributing to tumour behaviour. Many oncogenes and tumour-suppressor proteins, most prominently phosphatidylinositol-3-OH kinase (PI(3)K), MYC, RAS, p53, PTEN, $p16^{Ink4a}$ and retinoblastoma protein (RB), are frequent culprits in diverse cancers, although the overall mutational profiles of different cancer types can vary considerably. Tumour maintenance undoubtedly depends on the continued expression of certain oncogenes — a phenomenon known as oncogene addiction[6]. Lineage-dependency oncogenes that have key survival roles, in which genetic changes may be predetermined by the lineage programs inherent in the tumour precursor cell[7], are also likely to contribute. There is mounting evidence, however, that the nature of the cellular target has an important influence on tumour cell fate and

[1]Stem Cells and Cancer Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia. [2]Department of Medical Biology, The University of Melbourne, Parkville, Victoria 3010, Australia.

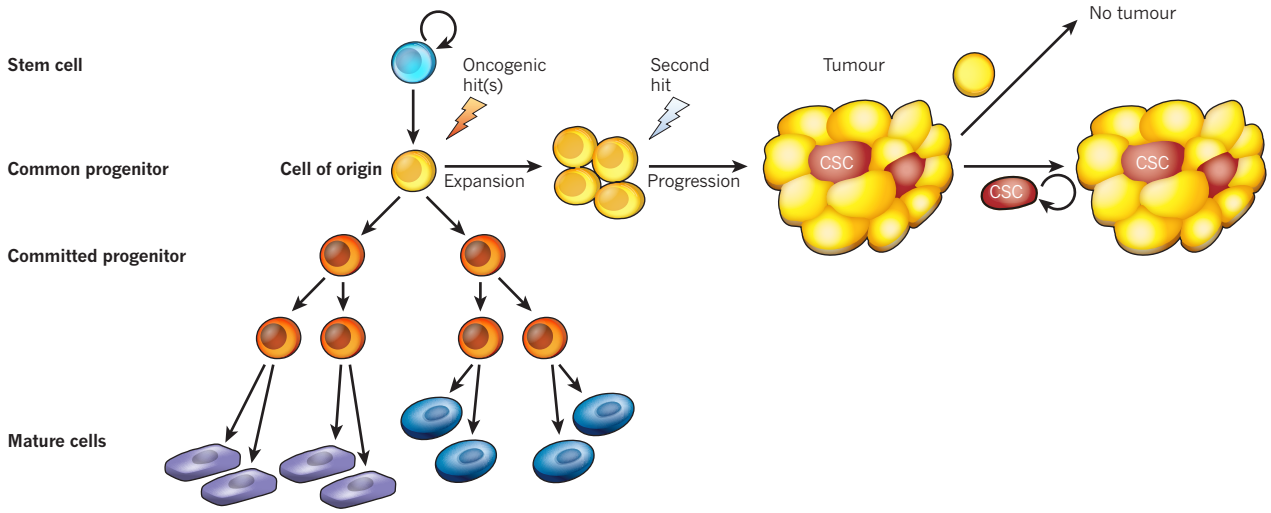


Figure 1 | The cell of origin and evolution of a cancer stem cell. Normal cellular hierarchy comprising stem cells that progressively generate common and more restricted progenitor cells, yielding all the mature cell types that constitute a particular tissue. Although the cell of origin for a particular tumour could be an early precursor cell such as a common progenitor, the accumulation of further epigenetic mutations by a cell within the aberrant population (in this case expanded) during neoplastic progression may result in the emergence of a CSC. In this model, only the CSCs (and not other tumour cells) are capable of sustaining tumorigenesis. Thus, the cell of origin, in which tumorigenesis is initiated, may be distinct from the CSC, which propagates the tumour.

pathology. Indeed, activation of the same oncogenic pathway in different cellular compartments or contexts may profoundly influence malignant potential[8]. For example, transgenic mouse models have shown that mutant *Hras* targeted to the hair follicle region highly predisposed mice to squamous carcinomas, whereas its targeting to more differentiated interfollicular or suprabasal cells resulted in papillomas with low malignant potential[9,10]. Moreover, transformation of distinct breast epithelial cells *in vitro* has indicated that the target cell is an important determinant of tumour phenotype[11].

Understanding the normal cellular hierarchy within a given tissue is an important prerequisite to identifying the cells of origin of cancers. Organ development proceeds in a hierarchical manner from stem cells to committed progenitor cells, which in turn yield differentiated cells that constitute the bulk of the tissue or organ (Fig. 1). The most primitive cells, stem cells, have been favoured candidates for targets of transformation because of their inherent capacity for self-renewal and their longevity, which would allow the sequential accumulation of genetic or epigenetic mutations required for oncogenesis. Nevertheless, any cell in the hierarchy with proliferative capacity could serve as a cell of origin in cancer, if it acquires mutations that re-instigate self-renewal capacity and prevent differentiation to a post-mitotic state.

The normal lineage hierarchy can serve as a framework to probe potential targets of carcinogenesis by comparison of lineage markers expressed on the surface of normal and neoplastic cell subsets. More accurate correlations, however, depend on comparisons of the expression signatures of normal cell populations with those of the different tumour subtypes arising within that organ. Notably, histologically indistinguishable glial-cell tumours from different parts of the central nervous system have distinct molecular gene signatures and chromosomal abnormalities, suggesting that they originated in different subpopulations of site-restricted progenitor cells[12,13]. In a recent integrated genomics approach to studying tumour heterogeneity, the transcriptomes of human brain tumours were matched to those of mouse neural stem cells (NSCs) from different cellular compartments within the central nervous system. Embryonic cerebral NSCs and adult spinal NSCs were revealed as the potential cells of origin for supratentorial and spinal ependymomas, respectively[14]. In breast cancer, the different molecular subtypes[15,16] have also been linked to normal epithelial subpopulations by the interrogation of specific gene-expression signatures[17]. Such observations remain correlative, however, until the tumorigenic potential of specific cells is proven *in vivo* by clonality or lineage-tracing studies. Although the hierarchy provides an important framework for understanding cells of origin in cancer, if tumour cells show phenotypic plasticity or dedifferentiate during neoplastic progression, then lineage markers and molecular signatures of tumour cells may not precisely reflect the true cell of origin in normal tissue.

Strategies to investigate the cellular origins of cancers

Genetically engineered mouse models have proven indispensable in addressing the cellular origin of cancers (Fig. 3). Two primary approaches have been used to tackle this question: one, transgenic or conditionally targeted gene technologies to explore the effects of oncogenes and tumour suppressors in different cellular contexts; and, two, genetic alteration of cells *ex vivo* before evaluating their tumorigenic capacity in mice. The first approach requires cell-specific promoters that direct expression of an oncogene, or Cre-mediated deletion of a tumour-suppressor gene, in a specific subset of cells *in vivo* (Fig. 3a). Ideally, such studies should use at least two promoters with different cell-type specificity to reveal the tumorigenic susceptibility of distinct cell subpopulations within that tissue. In this model, targeting of only one cell subpopulation is expected to reveal tumours that recapitulate the phenotype of the human cancer being modelled. Although this approach has been increasingly used to study cells of origin, particularly in brain tumours, it is often hampered by a lack of established cell-lineage-specific promoters, given that unique markers of stem and progenitor cells do not exist for the overwhelming majority of organs and tissues.

A further refinement of this *in vivo* targeting approach involves lineage tracing of cells as they undergo transformation. In this system, the main oncogenic event is activated conditionally in a limited number of cells rather than simultaneously in all cells that express the promoter. For example, a tamoxifen-inducible Cre recombinase–oestrogen receptor fusion protein (CreER)[18,19] driven by a cell-type-specific promoter allows inducible gene expression, in which the dose and number of pulses can be fine-tuned to ensure single-cell tracking. Lineage tracing at the clonal level is the current 'gold standard' for delineating the target cell of transformation in mouse models (Table 1).

In the second approach, defined cell subpopulations are genetically manipulated *ex vivo* and subsequently transplanted orthotopically into mice to assess their predisposition to tumour initiation (Fig. 3b). The strategy is applicable to cells from both human and mouse tissues, and

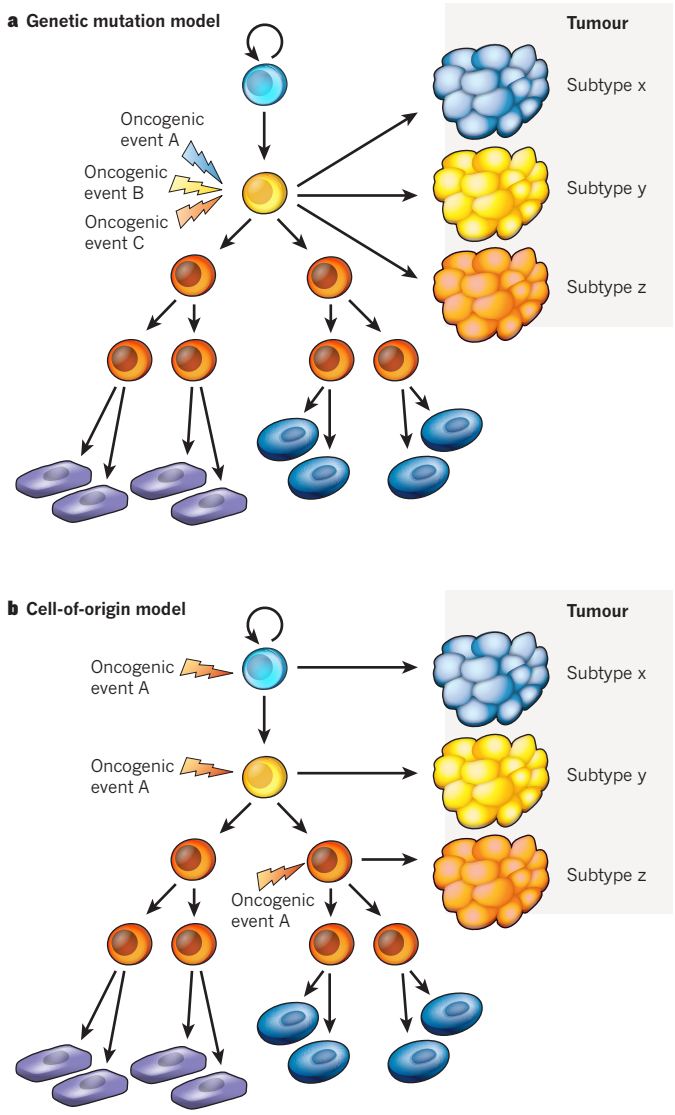


Figure 2 | Two models of intertumoral heterogeneity. **a**, In the genetic (and epigenetic) mutation model, mutations primarily determine the phenotype of the tumour, such that different mutations result in different tumour morphology. **b**, In the cell-of-origin model, different cell populations in the lineage hierarchy serve as cells of origin for the different cancer subtypes arising within that organ or tissue.

relies on the reproducible sorting of functionally defined populations that can serve as targets for the introduction of relevant oncogenic lesions. This approach has been widely exploited to identify potential cells of origin in human leukaemias, many of which contain characteristic chromosomal translocations.

An alternative way of exploring early cellular changes that occur before the onset of overt disease is to dissect the cellular components of preneoplastic tissue from individuals in families at high risk of cancer. These include carriers of germline mutations in the adenomatous polyposis coli (*APC*) gene, hereditary non-polyposis colorectal cancer (HNPCC) genes (such as *MSH2* and *MLH1*), and *BRCA1* or *BRCA2* genes. Carriers of mutated *APC* and HNPCC genes are predisposed to developing colorectal cancer[20]; female *BRCA1*- or *BRCA2*-mutation carriers are prone to breast and ovarian cancer[21]; and male *BRCA2*-mutation carriers often develop prostate cancer. This strategy has proven insightful in the case of *BRCA1*-mutation carriers (see 'Histopathology does not necessarily reflect cell of origin'). Combined with transplantation and clonality studies, **cell subsets predisposed to neoplastic progression can thus be identified.**

Cells of origin in haematopoietic malignancies

In different leukaemias, both normal stem and committed progenitor cells have been implicated as cellular targets of transformation. In chronic myeloid leukaemia (CML) — one of the first disorders to be defined by a dominant genetic mutation — the long-term haematopoietic stem cell (HSC) containing the *BCR–ABL* mutation has been established as the cell of origin by *in vivo* clonality studies in humans[22]. Although the HSC maintains the chronic phase of the disease, analysis of samples from patients in blast crisis — the acute and advanced stage of disease — has indicated that subsequent genetic events occurring in downstream progenitor cells give rise to leukaemia stem cells, highlighting the dynamic state of the tumorigenesis process[23]. The cells of origin for acute leukaemias, including myeloid, lymphoid and mixed-lineage, have not been definitively established. Human acute myeloid leukaemia (AML) may originate within the primitive haematopoietic cell compartment, on the basis of the similar cell-surface phenotypes of the leukaemia-initiating cell and the HSC, as well as lentivirus-mediated clonal-tracking studies[24]. A primitive human haematopoietic cell may also be the primary target of *MLL* fusion genes[25,26]. Moreover, *in vivo* evidence has implicated a human HSC-like cell as the initiating cell in a case of childhood leukaemia arising *in utero*[27].

Several studies have addressed potential cells of origin in mouse leukaemia models by transducing primary haematopoietic cell populations with oncogenes before transplantation, but these have yielded variable results. For mouse models of CML, only *BCR–ABL* targeted to HSCs, but not to committed progenitor cells, induced myeloproliferative disease[28], consistent with findings for human CML. Interestingly, the MLL–GAS7 fusion protein produced mixed lymphoid leukaemia when transduced into HSCs or multipotential progenitor cells but not when introduced into lineage-restricted progenitors. However, the MOZ–TIF2 (ref. 28), MLL–AF9 (ref. 29) and MLL–ENL[30,31] fusion proteins all initiated AML irrespective of the cell subtype transduced. Although HSCs generally appeared more susceptible to transformation than committed progenitors, a self-renewal program seemed to be reactivated in the latter cells during leukaemogenesis. In a 'knock-in' mouse model of *MLL–AF9*, only HSCs that expressed high levels of the fusion product and not the granulocyte–macrophage progenitors were transformed, but the latter could be efficiently transformed by a higher dose of *MLL–AF9* after retroviral transduction[32]. Thus, oncogene dosage affects transformation susceptibility, emphasizing the importance of using models that permit oncogene expression at levels relevant to human disease.

Further evidence that cancer can be initiated in cells other than stem cells has emerged from cell-fate mapping studies in transgenic mice overexpressing *Lmo2*: preleukaemic T-cell progenitors that had acquired self-renewal potential were identified as the cell of origin for T-cell acute lymphoblastic leukaemia (T-ALL) in this model[33]. Pertinently, mice lacking three pathways commonly repressed in cancer (p53, $p16^{Ink4a}$ and $p19^{Arf}$) contain cells that phenotypically resemble haematopoietic multipotential progenitor cells but have long-term reconstituting ability, indicating that they have acquired self-renewal capacity[34].

Cells of origin in solid tumours

Evidence is increasing that either stem or progenitor cells can act as targets for tumour initiation in a range of solid tumours (Table 1). Lineage-tracing studies (shown schematically in Fig. 4a) have identified probable cells of origin of intestinal, prostate and basal cell carcinomas, as well as pancreatic ductal adenocarcinoma, in mouse genetic models. Several other reports have used cell-specific promoters to drive Cre-mediated expression of the oncogenic event(s) in different cellular compartments of the mouse, whereas genetic manipulation of discrete cellular subsets has provided valuable insight into cell types prone to the initiation of carcinogenesis. It is crucial to note that although many studies have clearly identified the lineage in which the cancer originates, the precise cell type in the hierarchy (the cell of origin) in which transformation occurs remains elusive in most cases. Nevertheless, **in mouse models of intestinal and prostate tumours, it seems clear that the cancers**

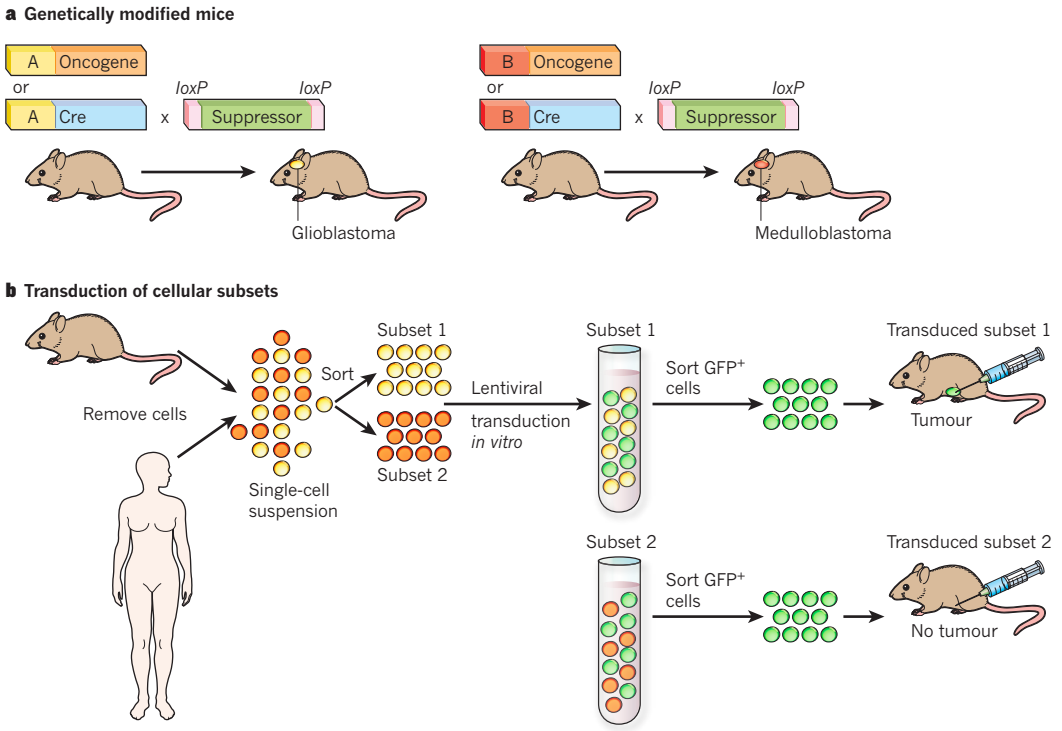


Figure 3 | Strategies used to identify cells of origin in cancer. a, Genetic mouse models can be used to either activate an oncogene or inactivate a tumour-suppressor gene in a discrete subpopulation of cells using cell-type-specific promoters. Comparative mouse models in which different promoters (A or B) drive expression of the same oncogenic lesion (either oncogene activation or tumour-suppressor inactivation) in the brain. Initiation of a glioblastoma is observed in the case of promoter A and a medulloblastoma in the case of promoter B, where promoters A and B are active in different cell subpopulations within the brain. **b**, Potential cells of origin in cancer can be addressed through the sorting of defined subpopulations from human or mouse tissue and their genetic manipulation *ex vivo*. Cell subpopulations are first transduced with genes encoding the oncogenic lesion(s), together with a fluorescent marker, then transplanted orthotopically into immunocompromised mice to evaluate the tumorigenic potency of the different subpopulations. GFP, green fluorescent protein.

originate in a bona fide stem cell that is capable of self-renewal and multilineage differentiation. Although the lineage in which the cancer originates has been revealed for skin, pancreatic, brain and breast tumours, the precise cells of origin have yet to be defined.

Two distinct crypt stem cells have been identified as the cells of origin for intestinal cancers. The vast majority of colorectal cancer is caused by mutations in the WNT signalling pathway, including loss of the negative regulator APC and activating mutations in β-catenin, both of which result in constitutive WNT activation[35]. *Apc* deletion in long-lived stem cells ($LGR5^+$) but not in short-lived transit-amplifying cells (using AhCre) revealed stem cells as a cell of origin for intestinal cancer in mice (Fig. 4b)[36]. This target cell is also marked by CD133 (also known as PROM1 or prominin 1)[37]. A novel intestinal stem cell located in the +4 or +5 position from the base of the crypt and therefore distinct from the $LGR5^+$ stem cell was also shown to be susceptible to tumorigenesis by deregulated WNT signalling using a BMI1–CreER knock-in model[38]. $LGR5^+$ stem cells are likely to be the target population for WNT-driven tumorigenesis in the stomach, where these cells seeded small adenomas[39]. The intestinal tumour load, however, precluded further lineage tracing of stem cells during the development of stomach cancer.

The cell of origin for brain cancers has been investigated using several mouse genetic models that have differed in design and the nature of the initiating oncogenic lesions. The models have predominantly included conditionally targeted mice, the RCAS–TVA system — in which gene transfer is mediated by an oncogene-carrying RCAS retrovirus to somatic cells in *TVA* transgenic mice[40] — and cell-culture-based analyses. Stereotactic injection of viruses into different areas of the brain[41,42] has also been used for the introduction of oncogenes or Cre recombinase to mimic focal tumorigenesis, but this approach cannot be used to identify cells of origin of cancer unequivocally as the transduced cell types are unknown. Although the available evidence argues for stem or multipotential neural progenitors in the subventricular zone (SVZ) as the primary cellular target for glioblastoma development, the cell of origin remains elusive owing to the complexity of this zone[43]. Many studies on brain tumorigenesis have used the nestin (*Nes*) or *Gfap* promoter regions to direct expression or inactivation: it is important to note that although both promoters drive expression in neural precursor cells, the *Gfap* promoter is also active in mature astrocytes[41]. Nonetheless, nestin-positive precursors were more susceptible to transformation by RAS and AKT than the GFAP-positive population, and produced high-grade glioblastomas, consistent with tumours originating in the stem/progenitor population[40]. Moreover, neural precursor cells in the SVZ of the adult brain efficiently initiated glioblastomas after conditional inactivation of *p53* (also known as *Trp53*), *Pten* and/or *Nf1* tumour-suppressor genes[41], and presymptomatic mice exhibited a premalignant cell population. By contrast, the more differentiated cell types in non-neurogenic areas of the adult brain proved less susceptible to malignant transformation[41,42]. Similarly, mice deficient in varying combinations of *p53*, *Pten* and/or *Rb* (also known as *pRb* and *Rb1*) developed tumours only in the SVZ and not from mature peripheral astrocytes[44]. Interestingly, the same stem/progenitor population seemed to initiate either gliomas or medulloblastomas, depending on the nature of the genetic lesions. More restricted progenitor cells may also initiate glioma development. Single-cell tracking studies of cells expressing mutant p53 implicate transit-amplifying cells in the SVZ[45]. Although oligodendrocyte progenitors and cells within the astrocyte compartment may also have the potential to seed glioblastomas[46,47], culturing cells before manipulation may not accurately reflect the *in vivo* situation, and the presence of more primitive cells within the cell cultures cannot be excluded. Thus, definitive evidence that mature astrocytes can serve as cells of origin for brain tumours awaits further experimentation.

Table 1 | Cells of origin (proven and candidate) identified in solid tumours by targeting distinct cellular subsets

Tumour type	Genetic model	Promoter–Cre construct	Lineage tracing	Cell of origin
Mouse models				
Brain: Glioblastoma	RAS, AKT activation (RCAS–TVA system: nestin, *Gfap* promoters)	NA	–	Neural progenitor cell[40]
	$p16^{Ink4a}$/$p19^{Arf}$, BMI1 inactivation; mutant EGFR	NA	–	Neural progenitor and astrocyte[46,47]
	p53, NF1 and/or PTEN inactivation	Nestin–CreERT2, Adeno–Cre	–	Multipotent progenitor[41]
	PDGFB activation (RCAS–TVA system)	NA	–	Oligodendrocyte progenitor[85]
	RAS, AKT activation; p53 inactivation	GFAP–Cre	–	Multipotent progenitor[42]
	Mutant p53 expression	GFAP–Cre	–	Neural progenitor or transit-amplifying cell[45]
	PTEN, p53 inactivation	GFAP–Cre	–	Multipotent progenitor[44]
Medulloblastoma	Patched inactivation	MATH1–Cre, GFAP–Cre	–	Multipotent progenitor and granule neuron progenitor[49]
	Smoothened activation	GFAP–Cre, MATH1–Cre, OLIG2–TVA–Cre	–	Multipotent progenitor and granule neuron progenitor[48]
	RB, p53, PTEN inactivation	GFAP–Cre	–	Multipotent progenitor[44]
	RB, p53 inactivation	Adeno–Cre	–	Neural progenitor cell[50]
	β-catenin mutant, p53 inactivation	BLBP–Cre, ATOH1–Cre	–	Dorsal brainstem progenitor[51]
Ependymoma (supratentorial)	$p16^{Ink4a}$/$p19^{Arf}$ inactivation; EPHB2 activation	NA	–	Embryonic cerebral stem/progenitor cell[14]
Intestine	APC inactivation	AhCre, LGR5–CreERT2	+	Stem cell[36]
	Mutant β-catenin	CD133–CreERT2	+	Stem cell[37]
	Mutant β-catenin	BMI1–CreER	+	Stem cell[38]
Lung	Kras activation	Adeno–Cre	–	Bronchioalveolar stem cell[77]
Mammary	NOTCH1 activation in cell subsets	NA	–	Luminal progenitor[65]
	BRCA1, p53 inactivation	BLG–Cre, K14–Cre	–	Luminal progenitor[63]
Pancreas	Kras activation, inflammation	RIP–CreER	+	Endocrine cell[69]
Prostate	PTEN inactivation	NKX3.1–CreERT2	+	Luminal stem cell[54]
	ERG1, PI(3)K and/or AR expression	NA	–	Basal progenitor[59]
	PTEN inactivation	PB–Cre	–	Basal progenitor[58]
	PTEN inactivation	PSA–Cre	–	Luminal cell[56,57]
Skin/basal cell carcinoma	Smoothened activation	K14–CreER	+	Interfollicular epidermal progenitor[72]
Stomach	APC inactivation	LGR5–CreERT2	+	Stem cell[39]
Human tissue				
Breast (basal-like subtype)*	Preneoplastic *BRCA1*$^{+/-}$ cell subsets	NA	–	Luminal progenitor[17]
Prostate	PI(3)K, ERG, AR into cell subsets	NA	–	Basal progenitor[60]

Adeno, adenoviral; Ah, cytochrome P450 1A1 gene (also known as *Cyp1a1*); AR, androgen receptor; BLG, β-lactoglobulin; K14, cytokeratin 14; NA, not applicable; PB, probasin (prostate-specific); PSA, prostate-specific antigen; RIP, rat insulin promoter.
*Refers to analysis of specific subsets from normal versus premalignant human breast tissue, leading to identification of a candidate cell of origin.

Unipotent cells within the mouse brain have been identified as the cell of origin for medulloblastoma. Constitutive hedgehog signalling (due to loss of *Ptc* or activation of *Smo*) in either the stem-cell compartment or granule neuron progenitor cells could initiate medulloblastomas but not astrocytomas or oligodendrogliomas[48,49]. Targeted deletion of a different set of genes (*p53* and *Rb*) also supports the notion that cerebellar stem cells or lineage-restricted granule progenitor cells can give rise to medulloblastomas[50]. Extending these observations further, it was shown that the cells must transition to the granule progenitor stage for the initiation of medulloblastomas, indicating that the true cell of origin for medulloblastomas that exhibit hedgehog pathway activation[48] is a unipotent progenitor cell[48]. A distinct cell type within the dorsal brainstem has recently emerged as the cell of origin for medulloblastomas that harbour activating mutations in the WNT pathway[51], indicating that different subtypes of medulloblastoma have distinct cellular origins. In mouse models of malignant peripheral nerve sheath tumours, tumours may initiate from differentiated glial cells in the adult brain[52,53].

Prostate cancer can originate from distinct cell lineages

Prostate cancers have been widely presumed to originate from mature luminal cells as these cancers are characterized by an expansion of luminal cells and the absence of basal cells. Recent findings, however, implicate distinct stem cells in the basal and the luminal cellular compartments, each of which can be targeted for oncogenesis by the loss of PTEN or PI(3)K activation. Rare luminal epithelial stem cells that express NKX3.1 and are castration resistant were identified[54], and these cells could initiate high-grade prostate intraepithelial neoplasia (PIN) and carcinomas. It is not yet clear whether these cells, termed CARNs, exist within 'normal' mouse or human prostate (that is, the

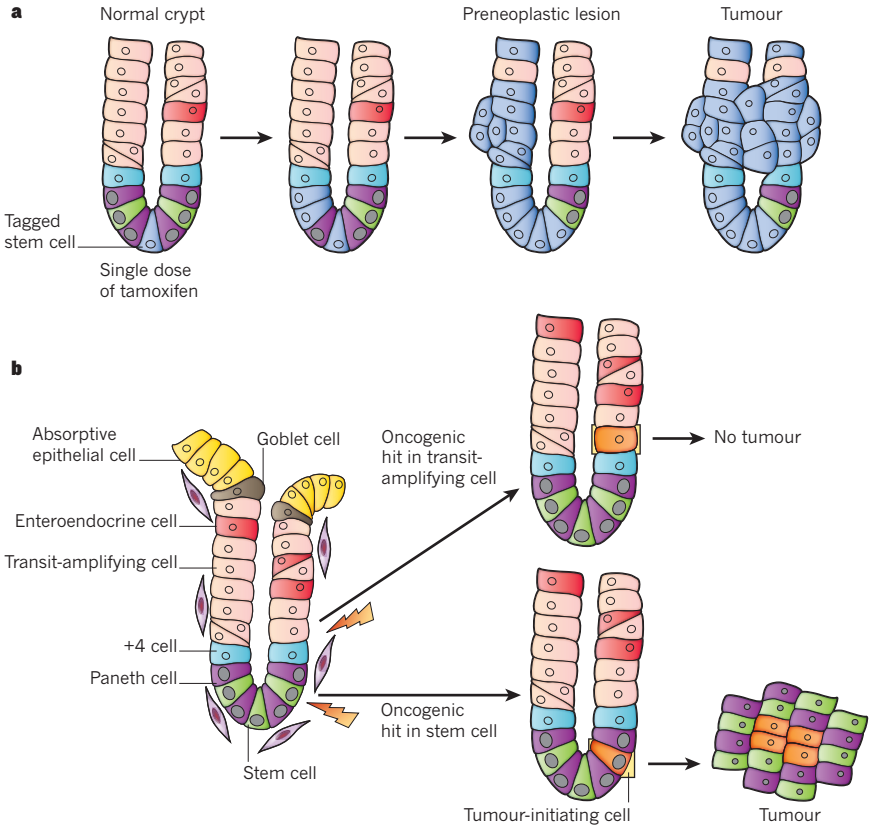


Figure 4 | Identification of crypt stem cells as the cell of origin in intestinal cancer by lineage tracing. **a**, Schematic depiction of lineage tracing in a colonic crypt, in which a single dose of tamoxifen can be used to activate CreER specifically in a stem cell to drive expression of a given oncogenic lesion in this cell population. The promoter that drives CreER expression will determine the cells in which this occurs. In the case of a stem cell, the incorporated reporter gene, such as *lacZ*, will mark all progeny of the stem cell. **b**, Schematic representation summarizing the data from ref. 36, in which lineage tracing of either stem or transit-amplifying cells deficient in *Apc* shows the initiation of intestinal tumours from stem cells only.

non-castrated state), but these bipotent, self-renewing cells may be mobilized as facultative stem cells during prostate regeneration after androgen withdrawal. Indeed, only the basal population isolated from normal mouse prostate has been demonstrated to contain stem cells with prostate-regenerative potential[55]. The question of whether CARNs exist in patients with prostate cancer is difficult to address as castration is implemented only after the development of advanced disease. The identification of a luminal cancer-initiating cell is consistent with findings that deletion of *Pten* in luminal cells of the mouse prostate leads to prostatic hyperplasia[56,57].

Conversely, basal cells have been demonstrated as an efficient target of tumorigenesis in a *Pten*-deficient mouse model[58] or when genetically manipulated *ex vivo* to overexpress *Erg*, androgen receptor (*Ar*) and/or PI(3)K, resulting in PIN lesions and carcinomas[59]. The tumorigenic susceptibility of purified basal and luminal subpopulations from human prostate tissue was recently evaluated [60]. When the cells were transduced with relevant oncogenic lesions, together with a fluorescent marker, and transplanted into immunocompromised mice, only the basal cells could initiate the development of prostate cancer reminiscent of the luminal-like cancers that arise in humans.

Histopathology does not necessarily reflect cell of origin

Tumours have largely been classified on the basis of their histological appearance and expression of markers (such as ER and HER2 in breast cancer) that predict the response of the tumour to a given treatment. However, the histological and cell-surface marker profiles of tumours do not necessarily predict the cell of origin, as illustrated above for prostate cancer. Other examples that underscore this point include breast, pancreatic and basal cell carcinomas, as discussed below.

Individuals that harbour mutations in the *BRCA1* tumour-suppressor gene develop breast cancers that usually resemble the basal-like subtype, typically associated with poor clinical prognosis[61,62]. The basal stem cell has therefore been presumed to be the transformation target for this tumour subtype, but the luminal progenitor has instead emerged as the likely cell of origin. Analysis of cellular subsets in precancerous breast tissue from *BRCA1*-mutation carriers demonstrated expansion of luminal progenitor cells that showed altered growth properties and aberrantly produced nodules when transplanted into mice (ref. 17 and F. Vaillant and J.E.V., unpublished data). Moreover, there are significant similarities between the gene-expression profiles of normal breast luminal progenitors, preneoplastic tissue from *BRCA1*-mutation carriers and basal-like breast cancers[17]. Indeed, inactivation of *Brca1* (and *p53*) in either luminal or basal cells of the mouse mammary gland showed that only the luminal cell population initiated basal-like cancers reminiscent of those arising in *BRCA1*-mutation carriers[63]. The presence of $ALDH1^+$ lobules in pathologically normal tissue from *BRCA1*-mutation carriers[64] is compatible with a luminal cell of origin, because luminal progenitors exhibit ALDH activity (F. Vaillant and J.E.V., unpublished data). NOTCH1 activation also targets luminal progenitor cells, generating an aberrant, self-renewing progenitor cell that yields mammary hyperplasia and, eventually, tumours[65]. Accordingly, high *NOTCH1* levels occur in basal-like breast cancers and predict poor prognosis[66]. The cells of origin for most other breast cancers have yet to be defined. In particular, the role of the mammary stem cell in breast oncogenesis is unclear, although WNT pathway activation primarily targets this population[67], and the 'claudin-low' subtype of breast cancer, which is characterized by low expression of genes involved in tight junctions and cell–cell adhesion, shares a similar molecular profile to that of the stem-cell subset[17,68].

Pancreatic ductal adenocarcinoma (PDAC) and premalignant ductal lesions (termed pancreatic intraepithelial neoplasia) have a ductal morphology, suggesting that they develop from pancreatic duct cells[69]. Unexpectedly, however, premalignant lesions were shown to derive from differentiated acinar cells that were reprogrammed to a duct-like phenotype[69–71]. Moreover, targeting a *Kras* oncogenic signal to insulin-positive endocrine cells induced PDAC. Notably, the ductal reprogramming of acinar cells required inflammatory tissue damage, highlighting a role for non-genetic factors in contributing to tumour phenotype. Ductal adenocarcinoma can also arise from other pancreatic cell lineages in the absence of tissue injury, for example PDX1-expressing cells[69].

Lineage-tracing studies have shown that basal cell carcinomas originate in progenitor cells resident in the interfollicular epidermis of the skin rather than from stem cells as originally postulated[72]. Conditional activation of hedgehog signalling in different cellular compartments, combined with cell-fate mapping, showed that long-lived progenitors in the interfollicular epithelium, but not the hair follicle bulge stem cell or the transit-amplifying cells, produced tumours. The block in differentiation evident in interfollicular epidermal cell clones with constitutive hedgehog signalling correlated with the expression of basal lineage markers (such as P-cadherin and keratins 7 and 15) and may have led to the notion that basal cell carcinomas arise from hair follicle bulge stem cells[8,73].

Potential relationships between cells of origin and CSCs

Although a stem cell may sustain the first oncogenic hit, subsequent alterations required for the genesis of a CSC can occur in descendent cells (Fig. 1). This is exemplified by CML, in which the HSC is the cell of origin in the more indolent phase of the disease but in patients with CML blast crisis, granulocyte–macrophage progenitors acquire self-renewal capacity through a β-catenin mutation and emerge as the probable CSC[23].

In some instances, particularly in early-stage cancers, the CSC may closely resemble the cell of origin, although this remains to be proven. For example, the leukaemia-initiating cell in AML[74] may prove to be the same as the leukaemia stem cell that propagates the disease. In a mouse model of intestinal cancer, despite all neoplastic cells arising from $CD133^+$ stem cells, only a small fraction of the tumour cells retained CD133 expression. It is tempting to speculate a hierarchical model of tumour progression in which this small subset of $CD133^+$ cells might generate the full repertoire of tumour cells and thereby correspond to CSCs. This notion is compatible with the observation that CD133 marks CSCs in certain human colorectal tumours[75,76]. Nevertheless, it remains to be determined whether these $CD133^+$ or $LGR5^+$ cells have tumour-propagating ability. Bronchioalveolar stem cells (BASCs) have been implicated as the cell of origin for lung adenocarcinomas induced by mutant *Kras*[77] in mice and may be closely related to the CSC, because the BASC marker Sca1 was recently shown to identify CSCs in certain mouse models of non-small-cell lung cancer[78]. In prostate cancer, if the oncogenic transformation of CARNs leads to the formation of CSCs in prostate cancer, then this might explain how early events occurring in the cell of origin can contribute to the emergence of hormone-refractory disease[54]. Although the relationships between tumour cells of origin and CSCs are not well understood, comprehensive cellular analyses of the preneoplastic and neoplastic states of different tumour subtypes should eventually shed light on this issue.

Therapeutic and diagnostic implications

Identification of the cell of origin has important implications for new preventive therapeutic approaches to suppress or reverse the initial phase of disease. Cancer chemoprevention will be most applicable to individuals within families at high risk of cancer such as *BRCA1/2*-mutation carriers. Cell-surface markers or proto-oncogenic kinases such as c-KIT[17] that show altered expression in cell subsets in preneoplastic tissue can be evaluated as prognostic markers, and for their ability to eradicate or modulate aberrant cells in either preneoplastic or established disease.

In principle, individuals that carry a defect in the *APC* gene, and are thus highly susceptible to colorectal cancer, could benefit from prophylactic treatment that targets APC-deficient cells for apoptosis. Tumour-necrosis-factor-related apoptosis-inducing ligand (TRAIL) in combination with all-*trans* retinoic acid selectively induced apoptosis in APC-deficient premalignant cells and intestinal polyps, thus inhibiting tumour growth[79]. Furthermore, treatment of biopsy samples of human colonic polyps from patients with familial adenomatous polyposis showed selective apoptosis of polyps, whereas normal tissue was unaffected, providing a potentially effective method of chemoprevention in these patients. In other families at high risk of colorectal cancer, with mutations in the *MLH1* or *MSH2* mismatch repair genes, inhibition (short-term and intermittent) of selective DNA polymerases may be a potential chemopreventive strategy, as these agents have been shown to elicit tumour cell death in patients with HNPCC[80].

Tamoxifen and aromatase inhibitors (inhibitors of oestrogen action and biosynthesis, respectively) are the prototypes for chemopreventive agents in hormone-receptor-positive breast cancer, as they markedly reduce the rate of disease recurrence and more than halve the incidence of new cancers in patients[81]. Recent findings have clarified how ovarian hormone exposure enhances breast cancer risk by showing that mammary stem cells, despite lacking receptors for these hormones, are highly responsive to steroid hormone signalling *in vivo*[82]. As the paracrine signals relayed to these stem cells seem to involve the receptor activator of NF-κB (RANK) signalling pathway, an exciting corollary of these findings is that it should be possible to prevent some forms of breast cancer by driving stem cells into a dormant state — for example, by blockade of the RANK pathway — for which inhibitors are already in clinical trial for bone metastases. There may also be prophylactic benefit for *BRCA1/2*-mutation carriers in the use of poly(ADP-ribose) polymerase (PARP) inhibitors[83], which are being evaluated for the treatment of *BRCA1/2*-associated breast cancers[84], if the early lesions in these individuals prove to be defective in DNA repair.

Perspective

It seems intuitive that both the cell of origin and the pattern of acquired mutations determine tumour fate and phenotype. The close association between cell lineage and cancer phenotype suggests that lineage-restricted mechanisms that normally operate during development may contribute to tumorigenesis. The cell of origin may often correspond to the normal tissue stem cell, exploiting its intrinsic self-renewal ability. This may particularly apply to tissues with very high turnover, such as the gut, because progenitor cells may not live long enough to acquire the full set of mutations required for malignancy. The stem cell or an early progenitor cell has also emerged as a likely cell of origin in certain leukaemias, glioblastomas and prostate cancer. In other malignancies, however, the initiating cell can be a restricted progenitor, as in the case of medulloblastomas, basal cell carcinomas and *BRCA1*-associated breast cancer. Indeed, in cell types that retain high proliferative potential, such as some differentiated lymphoid cells, the cell of origin could even be a mature cell type. Notably, there are several examples indicating that tumour phenotype may not directly reflect tumour histology or lineage marker expression, thus highlighting the requirement for *in vivo* studies to assess the propensity of cell populations to act as cells of origin.

Mouse models of oncogenesis have been pivotal in uncovering the cellular origins of cancer and the impact of specific mutations on tumorigenesis. Arguably, the choice of the genetically modified mouse model and the promoter/enhancer to recapitulate the effects of the oncogenic lesion has a major influence on tumour phenotype and behaviour. More specific promoters to drive expression of an initiating event within a definitive cellular compartment are likely to evolve as the normal lineage hierarchies within tissues are further refined. For studies on the cell of origin in human tissues, genetic and cellular analyses of tumour cell populations, at the single-cell level, from patients at different stages of disease should provide substantial insight into the relationships among normal cells, cells of origin and CSCs.

Identification of the cell of origin may permit a more systematic analysis of the genetic lesions involved in tumour initiation and progression, and serve as a platform for the identification of early disease biomarkers. It may also have important implications for preventing relapse, particularly in cases in which relapse results from a 'premalignant' clone (perhaps the cell of origin itself) that persists in the patient before acquiring a mutation that renders it malignant. If so, even patients with cancer who have a profound regression may require

maintenance therapy to reduce the chance of relapse. Finally, the gene signature of the cell of origin may elucidate key molecular pathways and driver mutations that could lead to new therapeutic approaches to prevent or target early-stage disease. ■

1. Visvader, J. E. & Lindeman, G. J. Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. *Nature Rev. Cancer* **8,** 755–768 (2008).
2. Marusyk, A. & Polyak, K. Tumor heterogeneity: causes and consequences. *Biochim. Biophys. Acta* **1805,** 105–117 (2010).
3. Ma, X. J. *et al.* Gene expression profiles of human breast cancer progression. *Proc. Natl Acad. Sci. USA* **100,** 5974–5979 (2003).
4. Weigelt, B. *et al.* Gene expression profiles of primary breast tumors maintained in distant metastases. *Proc. Natl Acad. Sci. USA* **100,** 15901–15905 (2003).
5. Tlsty, T. D. & Coussens, L. M. Tumor stroma and regulation of cancer development. *Annu. Rev. Pathol.* **1,** 119–150 (2006).
6. Weinstein, I. B. Addiction to oncogenes—the Achilles heal of cancer. *Science* **297,** 63–64 (2002).
7. Garraway, L. A. & Sellers, W. R. Lineage dependency and lineage-survival oncogenes in human cancer. *Nature Rev. Cancer* **6,** 593–602 (2006).
8. Perez-Losada, J. & Balmain, A. Stem-cell hierarchy in skin cancer. *Nature Rev. Cancer* **3,** 434–443 (2003).
9. Bailleul, B. *et al.* Skin hyperkeratosis and papilloma formation in transgenic mice expressing a *ras* oncogene from a suprabasal keratin promoter. *Cell* **62,** 697–708 (1990).
10. Brown, K., Strathdee, D., Bryson, S., Lambie, W. & Balmain, A. The malignant capacity of skin tumours induced by expression of a mutant H-*ras* transgene depends on the cell type targeted. *Curr. Biol.* **8,** 516–524 (1998).
11. Ince, T. A. *et al.* Transformation of different human breast epithelial cell types leads to distinct tumor phenotypes. *Cancer Cell* **12,** 160–170 (2007).
12. Sharma, M. K. *et al.* Distinct genetic signatures among pilocytic astrocytomas relate to their brain region origin. *Cancer Res.* **67,** 890–900 (2007).
13. Taylor, M. D. *et al.* Radial glia cells are candidate stem cells of ependymoma. *Cancer Cell* **8,** 323–335 (2005).
14. Johnson, R. A. *et al.* Cross-species genomics matches driver mutations and cell compartments to model ependymoma. *Nature* **466,** 632–636 (2010).
This paper demonstrates integrated genomic and cell-based approaches to identify cells of origin in ependymomas.
15. Sorlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98,** 10869–10874 (2001).
16. Sotiriou, C. *et al.* Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl Acad. Sci. USA* **100,** 10393–10398 (2003).
17. Lim, E. *et al.* Aberrant luminal progenitors as the candidate target population for basal tumor development in *BRCA1* mutation carriers. *Nature Med.* **15,** 907–913 (2009).
Identification of an aberrant cell population in preneoplastic tissue, and discovery that mutant-*BRCA1* tissue and basal cancers share a gene signature with normal luminal progenitors.
18. Hayashi, S. & McMahon, A. P. Efficient recombination in diverse tissues by a tamoxifen-inducible form of Cre: a tool for temporally regulated gene activation/inactivation in the mouse. *Dev. Biol.* **244,** 305–318 (2002).
19. Metzger, D. & Chambon, P. Site- and time-specific gene targeting in the mouse. *Methods* **24,** 71–80, (2001).
20. Hewish, M., Lord, C. J., Martin, S. A., Cunningham, D. & Ashworth, A. Mismatch repair deficient colorectal cancer in the era of personalized treatment. *Nature Rev. Clin. Oncol.* **7,** 197–208 (2010).
21. Narod, S. A. & Foulkes, W. D. *BRCA1* and *BRCA2*: 1994 and beyond. *Nature Rev. Cancer* **4,** 665–676 (2004).
22. Fialkow, P. J., Denman, A. M., Jacobson, R. J. & Lowenthal, M. N. Chronic myelocytic leukemia. Origin of some lymphocytes from leukemic stem cells. *J. Clin. Invest.* **62,** 815–823 (1978).
23. Jamieson, C. H. *et al.* Granulocyte–macrophage progenitors as candidate leukemic stem cells in blast-crisis CML. *N. Engl. J. Med.* **351,** 657–667 (2004).
The first functional evidence that cells of origin and cancer-propagating cells in a given malignancy are likely to be distinct.
24. Hope, K. J., Jin, L. & Dick, J. E. Acute myeloid leukemia originates from a hierarchy of leukemic stem cell classes that differ in self-renewal capacity. *Nature Immunol.* **5,** 738–743 (2004).
25. Barabe, F., Kennedy, J. A., Hope, K. J. & Dick, J. E. Modeling the initiation and progression of human acute leukemia in mice. *Science* **316,** 600–604 (2007).
26. Wei, J. *et al.* Microenvironment determines lineage fate in a human model of *MLL-AF9* leukemia. *Cancer Cell* **13,** 483–495 (2008).
27. Hong, D. *et al.* Initiating and cancer-propagating cells in *TEL-AML1*-associated childhood leukemia. *Science* **319,** 336–339 (2008).
28. Huntly, B. J. *et al.* MOZ-TIF2, but not BCR-ABL, confers properties of leukemic stem cells to committed murine hematopoietic progenitors. *Cancer Cell* **6,** 587–596 (2004).
29. Krivtsov, A. V. *et al.* Transformation from committed progenitor to leukaemia stem cell initiated by MLL–AF9. *Nature* **442,** 818–822 (2006).
30. Cozzio, A. *et al.* Similar MLL-associated leukemias arising from self-renewing stem cells and short-lived myeloid progenitors. *Genes Dev.* **17,** 3029–3035 (2003).
31. Drynan, L. F. *et al. Mll* fusions generated by Cre-*loxP*-mediated *de novo* translocations can induce lineage reassignment in tumorigenesis. *EMBO J.* **24,** 3136–3146 (2005).
32. Chen, W. *et al.* Malignant transformation initiated by *Mll-AF9*: gene dosage and critical target cells. *Cancer Cell* **13,** 432–440 (2008).
This study highlights the importance of gene dosage when assessing the effects of oncogene expression in candidate cells of origin.
33. McCormack, M. P. *et al.* The *Lmo2* oncogene initiates leukemia in mice by inducing thymocyte self-renewal. *Science* **327,** 879–883 (2010).
34. Akala, O. O. *et al.* Long-term haematopoietic reconstitution by $Trp53^{-/-}$ $p16^{Ink4a-/-}$ $p19^{Arf-/-}$ multipotent progenitors. *Nature* **453,** 228–232 (2008).
35. Clevers, H. Wnt/β-catenin signaling in development and disease. *Cell* **127,** 469–480 (2006).
36. Barker, N. *et al.* Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature* **457,** 608–611 (2009).
Using lineage-tracing studies, this paper demonstrates that colonic stem cells can act as the cell of origin for colon cancer. See also references 37 and 38.
37. Zhu, L. *et al.* Prominin 1 marks intestinal stem cells that are susceptible to neoplastic transformation. *Nature* **457,** 603–607 (2009).
38. Sangiorgi, E. & Capecchi, M. R. *Bmi1* is expressed *in vivo* in intestinal stem cells. *Nature Genet.* **40,** 915–920 (2008).
39. Barker, N. *et al.* $Lgr5^{+ve}$ stem cells drive self-renewal in the stomach and build long-lived gastric units *in vitro*. *Cell Stem Cell* **6,** 25–36 (2010).
40. Holland, E. C. *et al.* Combined activation of Ras and Akt in neural progenitors induces glioblastoma formation in mice. *Nature Genet.* **25,** 55–57 (2000).
41. Alcantara Llaguno, S. *et al.* Malignant astrocytomas originate from neural stem/progenitor cells in a somatic tumor suppressor mouse model. *Cancer Cell* **15,** 45–56 (2009).
42. Marumoto, T. *et al.* Development of a novel mouse glioma model using lentiviral vectors. *Nature Med.* **15,** 110–116 (2009).
43. Merkle, F. T., Mirzadeh, Z. & Alvarez-Buylla, A. Mosaic organization of neural stem cells in the adult brain. *Science* **317,** 381–384 (2007).
44. Jacques, T. S. *et al.* Combinations of genetic mutations in the adult neural stem cell compartment determine brain tumour phenotypes. *EMBO J.* **29,** 222–235 (2010).
45. Wang, Y. *et al.* Expression of mutant p53 proteins implicates a lineage relationship between neural stem cells and malignant astrocytic glioma in a murine model. *Cancer Cell* **15,** 514–526 (2009).
46. Bachoo, R. M. *et al.* Epidermal growth factor receptor and *Ink4a/Arf*: convergent mechanisms governing terminal differentiation and transformation along the neural stem cell to astrocyte axis. *Cancer Cell* **1,** 269–277 (2002).
47. Bruggeman, S. W. *et al.* Bmi1 controls tumor development in an Ink4a/Arf-independent manner in a mouse model for glioma. *Cancer Cell* **12,** 328–341 (2007).
48. Schuller, U. *et al.* Acquisition of granule neuron precursor identity is a critical determinant of progenitor cell competence to form Shh-induced medulloblastoma. *Cancer Cell* **14,** 123–134 (2008).
This study demonstrates that descendants of stem cells can act as crucial cellular targets of transformation. See also reference 49.
49. Yang, Z. J. *et al.* Medulloblastoma can be initiated by deletion of *Patched* in lineage-restricted progenitors or stem cells. *Cancer Cell* **14,** 135–145 (2008).
50. Sutter, R. *et al.* Cerebellar stem cells act as medulloblastoma-initiating cells in a mouse model and a neural stem cell signature characterizes a subset of human medulloblastomas. *Oncogene* **29,** 1845–1856 (2010).
51. Gibson, P. *et al.* Subtypes of medulloblastoma have distinct developmental origins. *Nature* **468,** 1095–1098 (2010).
52. Joseph, N. M. *et al.* The loss of *Nf1* transiently promotes self-renewal but not tumorigenesis by neural crest stem cells. *Cancer Cell* **13,** 129–140 (2008).
53. Zheng, H. *et al.* Induction of abnormal proliferation by nonmyelinating Schwann cells triggers neurofibroma formation. *Cancer Cell* **13,** 117–128 (2008).
54. Wang, X. *et al.* A luminal epithelial stem cell that is a cell of origin for prostate cancer. *Nature* **461,** 495–500 (2009).
This report establishes a new luminal stem cell as a target of prostate carcinogenesis and indicates a hierarchy of stem cells in this tissue.
55. Leong, K. G., Wang, B. E., Johnson, L. & Gao, W. Q. Generation of a prostate from a single adult stem cell. *Nature* **456,** 804–808 (2008).
56. Ma, X. *et al.* Targeted biallelic inactivation of *Pten* in the mouse prostate leads to prostate cancer accompanied by increased epithelial cell proliferation but not by reduced apoptosis. *Cancer Res.* **65,** 5730–5739 (2005).
57. Korsten, H., Ziel-van der Made, A., Ma, X., van der Kwast, T. & Trapman, J. Accumulating progenitor cells in the luminal epithelial cell layer are candidate tumor initiating cells in a *Pten* knockout mouse prostate cancer model. *PLoS ONE* **4,** e5662 (2009).
58. Mulholland, D. J. *et al.* $Lin^{-}Sca\text{-}1^{+}CD49f^{high}$ stem/progenitors are tumor-initiating cells in the *Pten*-null prostate cancer model. *Cancer Res.* **69,** 8555–8562 (2009).
59. Lawson, D. A. *et al.* Basal epithelial stem cells are efficient targets for prostate cancer initiation. *Proc. Natl Acad. Sci. USA* **107,** 2610–2615 (2010).
60. Goldstein, A. S., Huang, J., Guo, C., Garraway, I. P. & Witte, O. N. Identification of a cell-of-origin for human prostate cancer. *Science* **329,** 568–571 (2010).
This study demonstrates through transduction of isolated cell subsets that basal stem/progenitor cells are an important target cell.
61. Foulkes, W. D. *BRCA1* functions as a breast stem cell regulator. *J. Med. Genet.* **41,** 1–5 (2004).
62. Turner, N., Tutt, A. & Ashworth, A. Hallmarks of 'BRCAness' in sporadic cancers. *Nature Rev. Cancer* **4,** 814–819 (2004).
63. Molyneux, G. *et al. BRCA1* basal-like breast cancers originate from luminal

epithelial progenitors not from basal stem cells. *Cell Stem Cell* **7,** 403–417 (2010).
This study provides *in vivo* functional evidence that luminal progenitors rather than basal cells are an important target in the genesis of *BRCA1*-associated breast tumours.
64. Liu, S. *et al.* BRCA1 regulates human mammary stem/progenitor cell fate. *Proc. Natl Acad. Sci. USA* **105,** 1680–1685 (2008).
65. Bouras, T. *et al.* Notch signaling regulates mammary stem cell function and luminal cell-fate commitment. *Cell Stem Cell* **3,** 429–441 (2008).
66. Lee, C. W. *et al.* A functional *Notch–survivin* gene signature in basal breast cancer. *Breast Cancer Res.* **10,** R97 (2008).
67. Vaillant, F. *et al.* The mammary progenitor marker CD61/β3 integrin identifies cancer stem cells in mouse models of mammary tumorigenesis. *Cancer Res.* **68,** 7711–7717 (2008).
68. Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.* **12,** R68 (2010).
69. Gidekel Friedlander, S. Y. *et al.* Context-dependent transformation of adult pancreatic cells by oncogenic K-Ras. *Cancer Cell* **16,** 379–389 (2009).
70. De La, O. J. *et al.* Notch and Kras reprogram pancreatic acinar cells to ductal intraepithelial neoplasia. *Proc. Natl Acad. Sci. USA* **105,** 18907–18912 (2008).
71. Habbe, N. *et al.* Spontaneous induction of murine pancreatic intraepithelial neoplasia (mPanIN) by acinar cell targeting of oncogenic Kras in adult mice. *Proc. Natl Acad. Sci. USA* **105,** 18913–18918 (2008).
72. Youssef, K. K. *et al.* Identification of the cell lineage at the origin of basal cell carcinoma. *Nature Cell Biol.* **12,** 299–305 (2010).
73. Owens, D. M. & Watt, F. M. Contribution of stem cells and differentiated cells to epidermal tumours. *Nature Rev. Cancer* **3,** 444–451 (2003).
74. Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nature Med.* **3,** 730–737 (1997).
The first report indicating that early stem/progenitor cells are targeted for transformation in AML.
75. Ricci-Vitiani, L. *et al.* Identification and expansion of human colon-cancer-initiating cells. *Nature* **445,** 111–115 (2007).
76. O'Brien, C. A., Pollett, A., Gallinger, S. & Dick, J. E. A human colon cancer cell capable of initiating tumour growth in immunodeficient mice. *Nature* **445,** 106–110 (2007).
77. Kim, C. F. *et al.* Identification of bronchioalveolar stem cells in normal lung and lung cancer. *Cell* **121,** 823–835 (2005).
78. Curtis, S. J. *et al.* Primary tumor genotype is an important determinant in identification of lung cancer propagating cells. *Cell Stem Cell* **7,** 127–133 (2010).
79. Zhang, L. *et al.* Chemoprevention of colorectal cancer by targeting APC-deficient cells for apoptosis. *Nature* **464,** 1058–1061 (2010).
80. Martin, S. A. *et al.* DNA polymerases as potential therapeutic targets for cancers deficient in the DNA mismatch repair proteins MSH2 or MLH1. *Cancer Cell* **17,** 235–248 (2010).
81. Forbes, J. F. *et al.* Effect of anastrozole and tamoxifen as adjuvant treatment for early-stage breast cancer: 100-month analysis of the ATAC trial. *Lancet Oncol.* **9,** 45–53 (2008).
82. Asselin-Labat, M. L. *et al.* Control of mammary stem cell function by steroid hormone signalling. *Nature* **465,** 798–802 (2010).
83. Ashworth, A. Drug resistance caused by reversion mutation. *Cancer Res.* **68,** 10021–10023 (2008).
84. Tutt, A. *et al.* Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with *BRCA1* or *BRCA2* mutations and advanced breast cancer: a proof-of-concept trial. *Lancet* **376,** 235–244 (2010).
85. Lindberg, N., Kastemar, M., Olofsson, T., Smits, A. & Uhrbom, L. Oligodendrocyte progenitor cells can act as cell of origin for experimental glioma. *Oncogene* **28,** 2266–2275 (2009).

Acknowledgements I am grateful to J. Adams, G. Lindeman and A. Strasser for critical review of the manuscript, P. Dirks for discussion and P. Maltezos for preparation of figures. I apologize to authors whose work could not be cited owing to space limitations. J.E.V. is supported by the National Health and Medical Research Council and the Victorian Breast Cancer Research Consortium.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Correspondence should be addressed to the author (visvader@wehi.edu.au).

REVIEW

doi:10.1038/nature09782

Autophagy in immunity and inflammation

Beth Levine[1,2,3], Noboru Mizushima[4] & Herbert W. Virgin[5]

Autophagy is an essential, homeostatic process by which cells break down their own components. Perhaps the most primordial function of this lysosomal degradation pathway is adaptation to nutrient deprivation. However, in complex multicellular organisms, the core molecular machinery of autophagy — the 'autophagy proteins' — orchestrates diverse aspects of cellular and organismal responses to other dangerous stimuli such as infection. Recent developments reveal a crucial role for the autophagy pathway and proteins in immunity and inflammation. They balance the beneficial and detrimental effects of immunity and inflammation, and thereby may protect against infectious, autoimmune and inflammatory diseases.

There is only one known mechanism that eukaryotic cells possess to dispose of intracellular organelles and protein aggregates that are too large to be degraded by the proteasome. It is therefore not surprising that this mechanism — the lysosomal degradation pathway known as autophagy — is also used to degrade microorganisms (such as viruses, bacteria and protozoa) that invade intracellularly[1,2]. Indeed, the mutation of autophagy genes increases susceptibility to infection by intracellular pathogens in organisms ranging from plants to flies to worms to mice, and possibly to humans. Perhaps less apparent, but teleologically as intuitive, the autophagy pathway or unique functions of autophagy proteins also have a central role in controlling other diverse aspects of immunity in multicellular organisms.

The autophagy machinery is thought to have evolved as a stress response that allows unicellular eukaryotic organisms to survive during harsh conditions, probably by regulating energy homeostasis and/or by protein and organelle quality control. The same machinery might therefore be expected to diversify functionally in complex metazoan organisms, so as to regulate new layers of defences used by multicellular organisms to confront different forms of stress. A plethora of genetic, biochemistry, cell biology, systems biology and genomic studies have recently converged to support this notion. The autophagy machinery interfaces with most cellular stress-response pathways[3], including those involved in controlling immune responses and inflammation. This interface is not only at the level of the autophagy pathway, but also entails direct interactions between autophagy proteins and immune signalling molecules[4]. There is a complex reciprocal relationship between the autophagy pathway/proteins and immunity and inflammation; the autophagy proteins function in both the induction and suppression of immune and inflammatory responses, and immune and inflammatory signals function in both the induction and suppression of autophagy. Moreover, similar to cancer, neurodegenerative diseases and ageing[5], defects in autophagy — through autophagy gene mutation and/or microbial antagonism — may underlie the pathogenesis of many infectious diseases and inflammatory syndromes.

In this Review, we describe recent advances in our evolving comprehension of the interface between autophagy, immunity and inflammation. We discuss how emerging concepts about the functions of the autophagy pathway and the autophagy proteins may reshape our understanding of immunity and disease. This set of proteins not only orchestrates the lysosomal degradation of unwanted cargo, but also exerts intricate effects on the control of immunity and inflammation. Thus, the autophagy pathway and autophagy proteins may function as a central fulcrum that balances the beneficial and harmful effects of the host response to infection and other immunological stimuli.

Mechanisms and membrane dynamics of autophagy

Autophagy is a general term for pathways by which cytoplasmic material, including soluble macromolecules and organelles, is delivered to lysosomes for degradation[6]. There are at least three different types of autophagy, including macroautophagy, chaperone-mediated autophagy and microautophagy. Macroautophagy, usually referred to simply as autophagy, is the subject of this Review (Fig. 1). In this pathway, a portion of cytoplasm (usually 0.5–1 μm in diameter) is engulfed by an isolation membrane, or 'phagophore', resulting in the formation of a double-membrane structure known as the autophagosome. The outer membrane of the autophagosome fuses with the lysosome to become an autolysosome, leading to the degradation of autophagosomal contents by lysosomal enzymes. Autophagosomes can also fuse with endosomes or multivesicular bodies and major histocompatibility complex (MHC)-class-II-loading compartments[7]. Autolysosomes become larger as more autophagosomes and lysosomes fuse, but at a termination phase lysosomes are tubulated and fragmented for renewal[8].

The membrane dynamics of autophagosome formation involve complex processes, which are not completely understood. Nonetheless, the molecular dissection of autophagy membrane dynamics, stimulated by the discovery of *ATG* (autophagy-related) genes in yeast[9], has shed considerable light on this topic (Table 1). Several recent studies suggest that the endoplasmic reticulum (ER) is crucial for autophagosome formation. The ER cisternae often associate with developing autophagosomes, and electron tomography analysis has demonstrated direct connections between the ER and autophagosomal membranes[10,11]. Moreover, the function of several key autophagy proteins seems to be at the level of the ER (Fig. 1).

Autophagy is induced by nutrient starvation through the inhibition

[1]Department of Internal Medicine, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9113, USA. [2]Department of Microbiology, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9113, USA. [3]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9113, USA. [4]Department of Physiology and Cell Biology, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8519, Japan. [5]Department of Pathology and Immunology, Washington University School of Medicine and Midwest Regional Center of Excellence for Biodefense and Emerging Infectious Diseases Research, Campus Box 8118, 660 South Euclid Avenue, Saint Louis, Missouri 63110, USA.

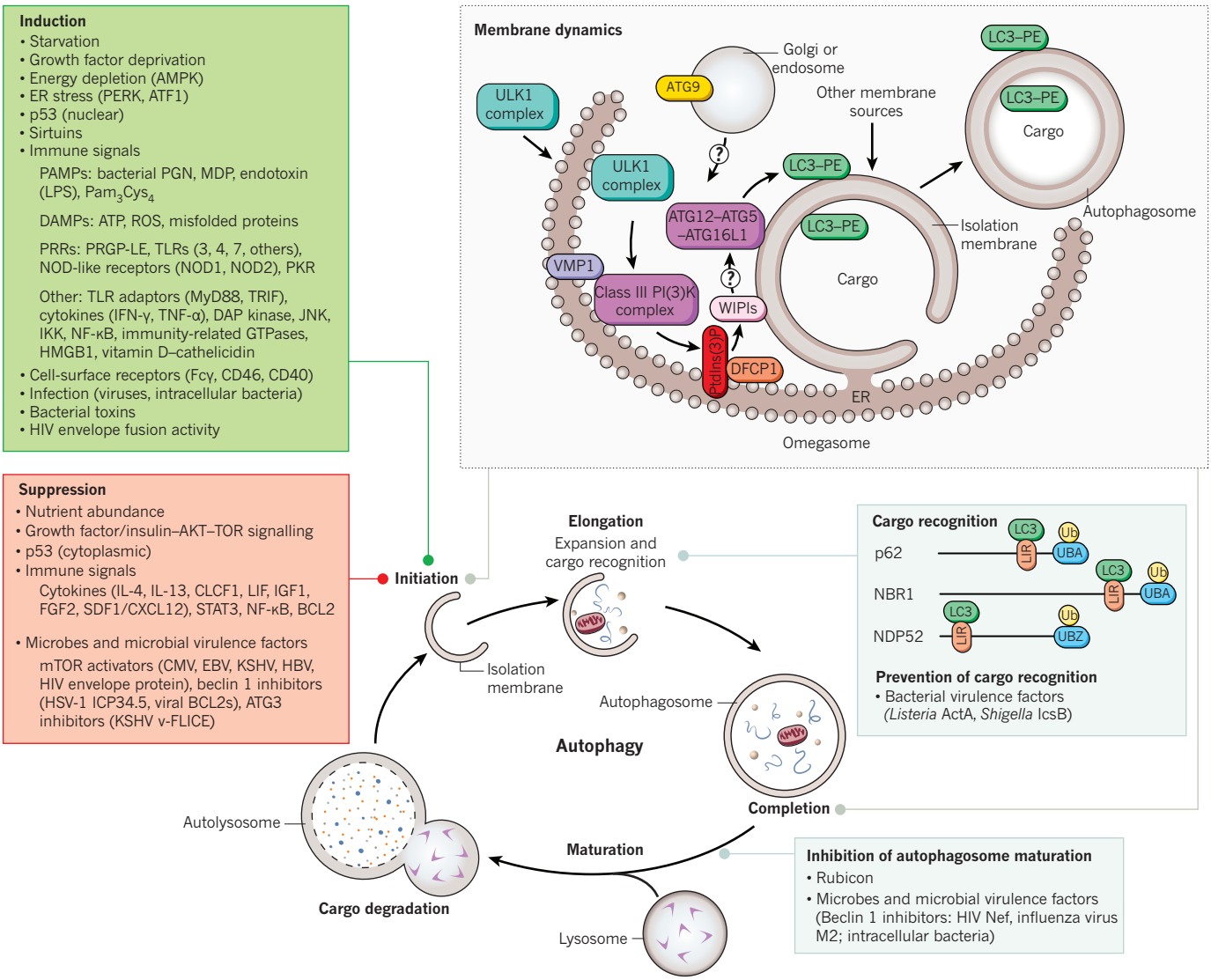


Figure 1 | Schematic overview of autophagy and its regulation. Overview of the autophagy pathway. The top right box shows a model of our current understanding of the molecular events involved in membrane initiation, elongation and completion of the autophagosome. The major membrane source is thought to be the endoplasmic reticulum (ER), although several other membrane sources, such as mitochondria and the plasma or nuclear membrane, may contribute. After induction of autophagy, the ULK1 complex (ULK1–ATG13–FIP200–ATG101) (downstream of the inhibitory mTOR signalling complex) translocates to the ER and transiently associates with VMP1, resulting in activation of the ER-localized autophagy-specific class III phosphatidylinositol-3-OH kinase (PI(3)K) complex, and the phosphatidylinositol-3-phosphate (PtdIns(3)P) formed on the ER membrane recruits DFCP1 and WIPIs. WIPIs and the ATG12–ATG5–ATG16L1 complex are present on the outer membrane, and LC3–PE is present on both the outer and inner membrane of the isolation membrane, which may emerge from subdomains of the ER termed omegasomes. The cellular events that occur during autophagy are depicted in the bottom diagram, including the major known cellular and microbial proteins that regulate autophagy initiation, cargo recognition and autophagosome maturation. Only those cellular proteins known to be adaptors for targeting microbes are shown; other proteins (not shown) also function in cargo recognition of mitochondria and other organelles. CMV, cytomegalovirus; DAMP, danger-associated molecular pattern; DAP, death-associated protein; EBV, Epstein–Barr virus; HBV, hepatitis B virus; HSV-1, herpes simplex virus 1; KSHV, Kaposi's sarcoma-associated herpesvirus; LIR, LC3-interacting region (motif); LPS, lipopolysaccharide; MDP, muramyl dipeptide; Pam_3Cys_4, a synthetic TLR2 agonist; PAMP, pathogen-associated molecular pattern; PERK, an eIF2α kinase; PGN, peptidoglycan; PRGP-LE, a peptidoglycan-recognition protein; PRR, pathogen-recognition receptor; ROS, reactive oxygen species; Ub, ubiquitin; UBA, ubiquitin-associated domain; UBZ, ubiquitin-binding zinc finger; v-FLICE, viral FLICE.

of mammalian target of rapamycin (mTOR), resulting in translocation of the mTOR substrate complex (ULK1/2, ATG13, FIP200 (also known as RB1CC1) and ATG101) from the cytosol to certain domains of the ER or closely attached structures[12,13]. This leads to the recruitment of the class III phosphatidylinositol-3-OH kinase (PI(3)K) complex, which includes at least VPS34 (also known as PIK3C3), VPS15 (PIK3R4 and p150), beclin 1 and ATG14, to the ER[13,14]. The PI(3)K complex produces phosphatidylinositol-3-phosphate (PtdIns(3)P), which recruits effectors such as double FYVE-containing protein 1 (DFCP1) and WD-repeat domain phosphoinositide-interacting (WIPI) family proteins. DFCP1 is diffusely present on the ER or the Golgi, but translocates to the autophagosome formation site in a PtdIns(3)P-dependent manner to generate ER-associated Ω-like structures termed omegasomes[15]. Among the four WIPI isoforms, WIPI2 is the major form in most cell types and functions downstream of DFCP1, and it may promote the development of omegasomes into isolation membranes or autophagosomes[16].

An ER-associated multispanning membrane protein, VMP1, is also important for autophagosome formation. Although VMP1 interacts with beclin 1 and is present at the autophagosome formation site at an early stage, it seems to function at a late stage in autophagy[13,17,18].

Table 1 | Key proteins involved in mammalian autophagosome formation and their immune functions

Protein complex	Function of protein complex in autophagy	Specific protein	General properties	Immunological/host defence functions
Nucleation step of autophagosome formation				
ULK complex	This complex is negatively regulated by mTORC1 in a nutrient-dependent manner. After autophagy induction, this complex translocates to early autophagic structures. Although FIP200 and ATG13 are known to be phosphorylated by ULK1, physiologically relevant substrates remain unknown. FIP200 and ATG101 may have functions similar to yeast Atg17, 29 and 31, although they show no sequence similarity with these proteins.	ULK1/2	Protein kinase, phosphorylated by mTORC1	Antibacterial[47*,48*]; antiviral[46*]
		ATG13	Phosphorylated by mTORC1	Unknown
		FIP200	Scaffold for ULK1/2 and ATG13	Maintains numbers of fetal haematopoietic stem cells[72]
		ATG101	Interacts with ATG13	Unknown
Class III PI(3)K complex	Beclin 1 is negatively regulated by BCL2 and by BCL-X_L through direct binding. This complex produces PtdIns(3)P, probably on the ER. VPS34, VPS15 and beclin 1 are shared with the UVRAG complex, which seems to function in the late endocytic pathway. Rubicon negatively regulates autophagosome–lysosome fusion through interaction with the UVRAG complex.	VPS34	PI(3) kinase	Antiviral[45*]; phagosome maturation[33]
		VPS15	Myristoylated	Unknown
		Beclin 1	BH3-only protein, interacts with BCL2 and BCL-X_L	Antibacterial[45*,48*]; antiviral[46*,57,58]; apoptotic corpse clearance[84]; decreases inflammation in tumours[5]; regulates germinal centre induction[5]; phagosome maturation[31,33,49]; interacts with TLR signalling adaptors[3]
		ATG14	Autophagy-specific subunit	Unknown
		AMBRA1	Interacts with and activates beclin 1	Unknown
		UVRAG	A VPS38 homologue; interacts with class C VPS (HOPS) complex	Unknown
		Rubicon	Interacts with beclin 1	Unknown
Others	DFCP1 forms an omegasome on the ER, at which other ATG proteins are assembled. ATG9, WIPIs and VMP1 are present on the autophagic membrane. ATG9 also exists in other compartments such as endosomes and the Golgi apparatus.	ATG2	Interacts with Atg18 in yeast	Antiviral[46*]
		ATG9	Transmembrane protein	Antiviral[46*]; inhibits innate immune signalling[4]
		WIPI1–4	PtdIns(3)P-binding proteins	Unknown
		DFCP1	PtdIns(3)P-binding ER protein	Unknown
		VMP1	Transmembrane protein	Unknown
Elongation step				
ATG12-conjugation system	The ATG12–ATG5–ATG16L1 dimer is important for LC3–PE conjugation. This complex is present on the outer side of the isolation membrane and is essential for proper elongation of the isolation membrane.	ATG12	Ubiquitin-like, conjugates to ATG5	Antiviral[46*]; antibacterial[34]; antigen presentation[7,32,74]; inhibits type I IFN production[78]
		ATG7	E1-like enzyme	Antiviral[45*,46*]; antibacterial[48*]; antigen presentation[32]; phagosome maturation[31]; maintains number of T cells[44,72]; intestinal immune epithelial cell function[90]; inhibits type I IFN production[78]; inhibits pro-inflammatory cytokine production[82]
		ATG10	E2-like enzyme	Unknown
		ATG5	Conjugated by ATG12	Antiviral[40,46*]; antibacterial[1,30,35,47*,100]; antiparasitic[35]; antigen presentation[27,32,73]; phagosome maturation[31]; apoptotic corpse clearance[84]; maintains number of T cells[44,72]; maintains number of B1a B cells[44]; intestinal immune epithelial cell function[90]; inhibits type I IFN production[77,78]; increases type I IFN production[76]
		ATG16L1	Homodimer, interacts with ATG5	Antibacterial[87,88]; antigen presentation[52]; Crohn's disease risk allele[85]; intestinal immune epithelial cell function[81,90]; inhibits pro-inflammatory cytokine production[82]
LC3-conjugation system	The formation of LC3–PE conjugates and their deconjugation by ATG4 is important for isolation membrane elongation and/or complete closure. LC3 is present on both the inner and outer membrane of the autophagosomes, and also serves as an adaptor for selective substrates such as p62, NBR1, NDP52 and the yeast mitophagy protein Atg32.	LC3 (GATE16, GABARAP)	Ubiquitin-like, conjugates to PE	Antiviral[46*]; antibacterial[48*]; antigen presentation[7]; adaptor for substrates of selective autophagy of microbes (xenophagy)[38,41,61]
		ATG4A–D	LC3 carboxy-terminal hydrolase, deconjugating enzyme	Antiviral[46*]
		ATG7	E1-like enzyme	Antiviral[45*,46*]; antibacterial[48*]; antigen presentation[32]; phagosome maturation[31]; maintains number of T cells[44,72]; intestinal immune epithelial cell function[90]; inhibits type I IFN production[78]; inhibits pro-inflammatory cytokine production[82]
		ATG3	E2-like enzyme	Antiviral[45*]

The components of the autophagy machinery that have been shown to participate in the immune and inflammatory processes depicted in Fig. 3. Owing to space limitations, primary papers are cited only for those citations also included in the main text; otherwise, see references contained within cited review articles.
*Function observed in model organism (for example, *Dictyostelium discoideum*, *Nicotiana benthamiana*, *Arabidopsis thaliana*, *Drosophila melanogaster* or *Caenorhabditis elegans*).

This is perhaps consistent with other evidence that beclin 1–class III PI(3)K complexes may function in autophagosomal maturation (in addition to vesicle nucleation), a process that can be regulated by other beclin-1-interacting partners such as rubicon (Table 1). At the final step of autophagosome formation, elongation of the isolation membrane and/or completion of enclosure require two ubiquitin-like conjugates. The first is the ATG12–ATG5 conjugate, which is produced by the ATG7 (E1-like) and ATG10 (E2-like) enzymes, and functions as a dimeric complex together with ATG16L1 (ref. 19) . The second is the phosphatidylethanolamine (PE)-conjugated ATG8 homologues — LC3, GATE16 and GABARAP — which are produced by the ATG7 and ATG3 (E2-like) enzymes[9,20]. Although the proteins involved in autophagosome membrane formation have been characterized as discrete complexes (Table 1), several potential interconnections between components of the different complexes were identified by a recent proteomics study[21]. Such interconnections may function in autophagosome membrane formation or other distinct cellular functions. For example, the ATG12–ATG3 conjugate is implicated in mitochondrial homeostasis but not in autophagosome membrane formation[22].

In addition to the ER, other membranes may be involved in autophagosome formation (Fig. 1). ATG9, another multispanning membrane protein, is essential for autophagy[23] and traffics between the *trans*-Golgi network, endosomes and autophagosome precursors[24]. Studies suggest that mitochondria, the plasma membrane and the nuclear membrane could also be membrane sources for autophagosome formation[25–27]. However, the lack of detection of specific protein markers for these structures on the autophagosomal membrane leaves the decades-old question of the membrane source of the autophagosome unanswered. It is possible that cells may use different membrane sources to form the autophagosome in different contexts, thereby permitting specialization of membrane dynamics in a manner that allows divergent autophagy-inducing signals to stimulate the capture of spatially distinct cargo.

Selective autophagy tackles microbes

Autophagy was originally considered to be a non-selective bulk degradation process, but it is now clear that autophagosomes can degrade substrates in a selective manner[28]. In addition to endogenous substrates, autophagy degrades intracellular pathogens in a selective form of autophagy, termed xenophagy. Similar to bulk autophagy (such as that induced by nutrient deprivation) and other forms of selective autophagy (such as degradation of damaged mitochondria, peroxisomes, aggregate-prone proteins or damaged ER), the precise membrane dynamics and specificity determinants of xenophagy are not fully understood. Nonetheless, considerable advances have been made, and interesting similarities and differences are beginning to emerge between cellular recognition and degradation of self versus foreign microbial components through autophagy-like pathways (Figs 1 and 2).

The vacuoles used for the engulfment of intracytoplasmic bacteria are similar to autophagosomes, and their formation requires the core autophagy machinery. But one apparent difference is the vacuole size; for example, the diameter of group A *Streptococcus*-containing autophagosome-like vacuoles (GcAV) can be as big as 10 μm. These large GcAVs are generated by the RAB7-dependent fusion of small isolation membranes[29]. By contrast, the formation of starvation-induced autophagosomes requires RAB7 later in the autophagy process, at the autophagosome–lysosome fusion step.

A more complex question is how autophagosomes (or components of the autophagy pathway) capture pathogens that are inside vacuolar compartments (Fig. 2). There are at least four general pathways that may be used for autophagy-protein-dependent targeting of bacteria to the lysosome. These include autophagy-protein-facilitated fusion of bacteria-containing phagosomes with lysosomes, the envelopment of bacteria-containing phagosomes or endosomes by autophagosomal membranes, the fusion of bacteria-containing phagosomes or endosomes with autophagosomes, or the xenophagic capture of bacteria that have escaped inside the cytoplasm. In some cases, the route of autophagy-dependent targeting to the lysosome has been well defined, such as for group A *Streptococcus* that escapes from endosomes[30]. For several bacteria, however, the precise route is unclear. Many studies define bacterial autophagy as the co-localization of bacteria and LC3, but we now know that LC3 can decorate membranous compartments other than autophagosomes (including phagosomes).

Several lines of evidence suggest that autophagy proteins function more broadly, not only in classical macroautophagy, but also in the process of phagolysosomal maturation during antigen presentation and microbial invasion. Autophagy proteins are required for the fusion of phagosomes that contain Toll-like receptor (TLR)-ligand-enveloped particles with lysosomes in macrophages[31], and for the fusion of phagosomes that contain TLR-agonist-associated apoptotic cell antigens with lysosomes in dendritic cells during MHC class II antigen presentation[32]. The self ligand and cell-surface receptor SLAM functions as a microbial sensor that recruits the beclin 1–class III PI(3)K complex to phagosomes containing Gram-negative bacteria, facilitating phagolysosomal fusion and activation of the antibacterial NADPH oxidase (NOX2) complex[33]. In addition, the engagement of TLR or Fcγ receptors during phagocytosis recruits LC3 (and ATG12) to the phagosome through NOX2-dependent generation of reactive oxygen species (ROS)[34]. Thus, in bacterial infections, a paradigm is emerging in which the coordinated regulation of microbial sensing, phagolysosomal maturation and antibacterial activity involves the recruitment of autophagy proteins to the phagosome. As a corollary, an interesting speculation is that impaired recruitment of autophagy proteins to the phagosome may contribute to the pathogenesis of chronic granulomatous disease, a genetic disorder caused by mutations in the *NOX2* gene (also known as *CYBB*) and characterized by recurrent bacterial and fungal infections and inflammatory complications, such as inflammatory bowel disease.

Another autophagosome-independent function of autophagy proteins in pathogen destruction has been described in interferon-γ (IFN-γ)-treated macrophages infected with the parasite *Toxoplasma gondii*. The parasite-derived membrane, termed the parasitophorous vacuole, undergoes destruction through a mechanism that involves ATG5-dependent recruitment of the immunity-related GTPase proteins to the parasitophorous vacuole[35,36], leading to the death of the parasite in the infected cell[35,37]. Together, these studies suggest that autophagy proteins have diverse roles in membrane dynamics to benefit the host in the removal of invading pathogens (Fig. 2), through xenophagy, phagolysosomal maturation, the recruitment of molecules that damage pathogen-derived membranes, and presumably, many other as yet undiscovered mechanisms.

The mechanisms that cells use to target intracellular bacteria (and probably viruses) to autophagosomal compartments are notably similar to those used for selective autophagy of endogenous cargo. Cellular cargo is commonly targeted to autophagosomes by interactions between a molecular tag (such as polyubiquitin), adaptor proteins such as p62 (also known as SQSTM1 or sequestome 1) or NBR1 (which recognize these tags and contain an LC3-interacting region (LIR) characterized by a WXXL or WXXI motif), and LC3 (ref. 28). These adaptor molecules enable autophagy to target designated cargo selectively to nascent LC3-positive isolation membranes. As reviewed elsewhere[38], a similar mechanism involving ubiquitin and p62 seems to be involved in the targeting of intracellular bacteria, such as *Salmonella enterica* serotype Typhimurium (*S.* Typhimurium), *Shigella flexneri* and *Listeria monocytogenes*, to autophagosomes.

After escape into the cytoplasm or in vacuolar membrane compartments damaged by type III secretion system (T3SS) effectors, bacteria or bacteria-containing compartments, respectively, may become coated with ubiquitin and associate with p62 and nascent LC3-positive isolation membranes. The autophagosomal targeting of *Salmonella* also requires another cellular factor, NDP52 (nuclear dot protein 52), an autophagy adaptor protein that, like p62, contains an LIR and ubiquitin-binding domains and restricts intracellular bacterial replication. A ubiquitin-independent pathway (that does not involve p62 or NDP52) could also function in targeting damaged *Salmonella*-containing vacuoles (SCVs)

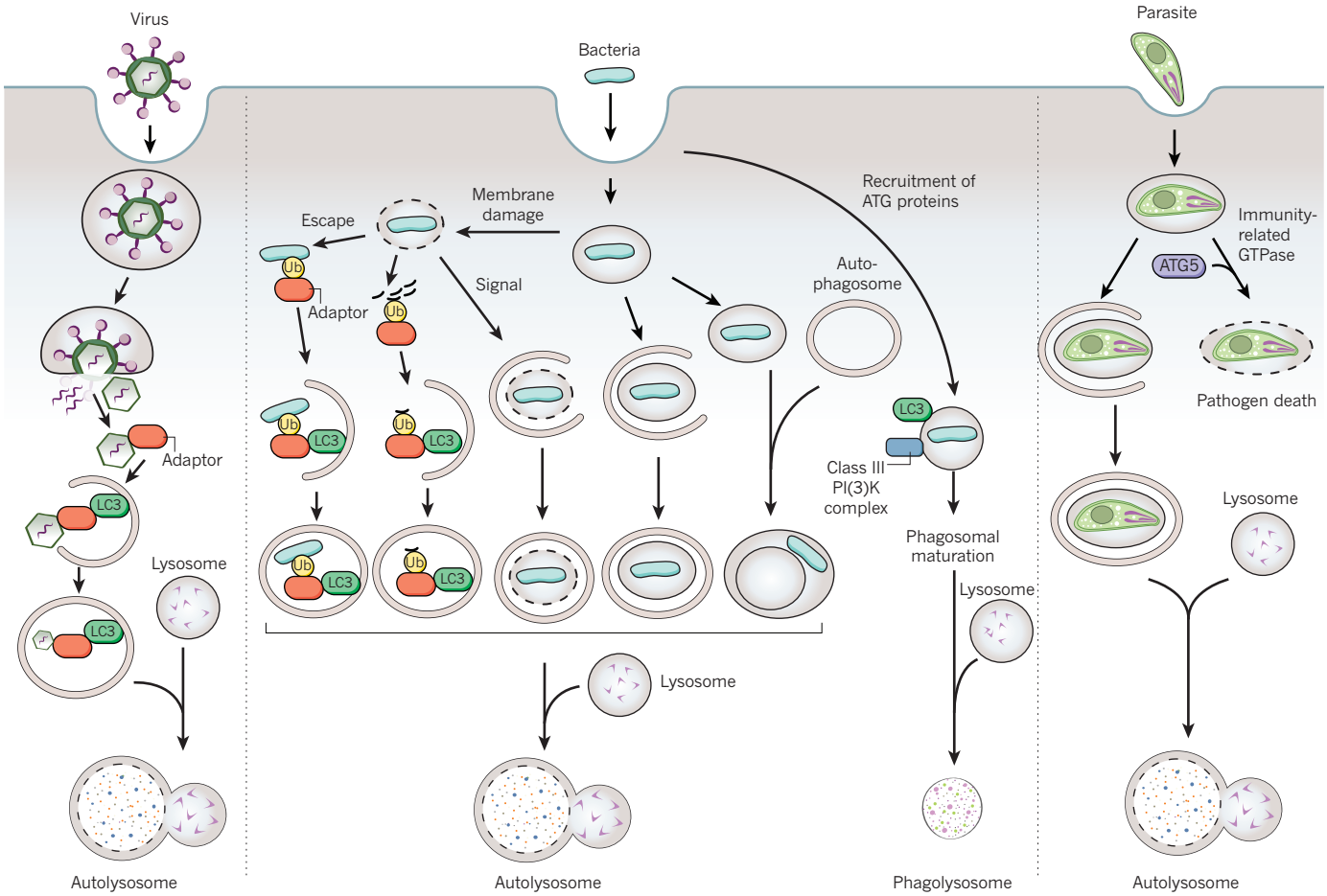


Figure 2 | Possible autophagy-protein-dependent pathways of pathogen degradation. Possible pathways involving the autophagy machinery by which viruses, bacteria (and damaged membranes of bacteria-containing vacuoles) and parasites may be targeted to the lysosome. Adaptor refers to the proteins shown in the cargo-recognition box in Fig. 1; however, as yet undiscovered adaptors may be involved in pathogen recognition, and pathogen targeting may involve ubiquitin-dependent or -independent mechanisms.

to the autophagosome. In this pathway, a lipid second messenger, diacylglycerol, acts as a signal for the co-localization of SCVs with LC3-positive autophagosomes by a mechanism that involves protein kinase C and its downstream targets, JNK and NADPH oxidase[39]. The autophagic targeting of a cytoplasmic positive-strand RNA virus, Sindbis virus, also occurs in a ubiquitin-independent manner, but involves the interaction of p62 with the viral capsid protein[40]. Thus, diverse molecular strategies, including ubiquitin-dependent and -independent mechanisms, may be used to target microbes inside the cytoplasm or vacuolar compartments to the autophagosome.

Beyond targeting intracellular pathogens for degradation, p62 may have further beneficial effects in infected host cells. For example, *Shigella* vacuolar membrane remnants generated by bacterial T3SS-dependent membrane damage are targeted by polyubiquitination, p62 and LC3 for autophagosomal degradation[41] (Fig. 2). These membrane remnants also accumulate numerous molecules involved in sensing and transduction of pathogen-associated molecular pattern (PAMP) and danger-associated molecular pattern (DAMP) signals, and there is an increase in nuclear factor-κB (NF-κB)-dependent cytokine production, ROS production and necrotic cell death in autophagy-deficient cells. Thus, the ubiquitin–p62-dependent autophagic targeting of pathogen-damaged membranes could help to control detrimental downstream inflammatory signalling during bacterial invasion into host cells. Another emerging concept is that selective autophagy of viral proteins, similar to selective autophagy of aggregate-prone toxic cellular proteins, may protect post-mitotic cells such as neurons against cell death. For example, in Sindbis-virus-infected mice with neuron-specific inactivation of Atg5, there is an accumulation of Sindbis virus antigens (without increased levels of infectious virus), increased neuronal cell death and increased animal mortality[40]. Moreover, p62 is required for starvation and IFN-γ-induced targeting of Fau (and perhaps other ubiquitylated protein complexes) to mycobacteria-containing phagosomes, resulting in the generation of antimycobacterial Fau-derived peptides[42].The role of p62 in innate immunity is probably evolutionarily ancient, as the *Drosophila* p62 orthologue REF(2)P was originally identified in a screen for modifiers of sigma virus replication[43].

We speculate that p62, as well as the other known LC3-interacting adaptor proteins (NBR1 and NDP52), may represent the tip of the iceberg in terms of cellular adaptor proteins that bind to ubiquitin (or other molecular tags) and target microbial substrates and cytosolic material to autophagosomes to coordinate innate immune responses. A recent proteomics study showed that the mammalian ATG8 family, which includes LC3, GATE16 and GABARAP, has 67 high-confidence interactions with other cellular proteins[21]. Some of these new ATG8-family-member-interacting partners may have an as yet undiscovered role in innate immunity. Another open question is whether the known proteins involved in selective autophagy of mitochondria (called mitophagy), such as Nix (also known as BNIP3L) and parkin[44], also function in microbial autophagy.

Autophagy and resistance to infection

The autophagy pathway and/or autophagy proteins have a crucial role in resistance to bacterial, viral and protozoan infection in metazoan organisms. The genetic deletion or knockdown of autophagy genes

protects plants from viral, fungal and bacterial infection by preventing the uncontrolled spread of programmed cell death during the plant innate immune or hypersensitive response[45]. In other organisms, autophagy proteins function in a cell-autonomous manner to control infection by intracellular pathogens. In *Drosophila*, autophagy gene mutation increases susceptibility to viral (vesicular stomatitis virus)[46] and bacterial (*L. monocytogenes*)[47] infection. In *Dictyostelium* and *Caenorhabditis elegans*, autophagy gene mutation increases susceptibility to lethal *S.* Typhimurium infection[48]. In mice, knockout of Atg5 in macrophages and neutrophils increases susceptibility to infection with *L. monocytogenes* and the protozoan *T. gondii*[35], and neuron-specific Atg5 knockout increases susceptibility to central nervous system Sindbis virus infection[40]. As noted in the next section, the autophagy pathway and proteins may also have 'proviral' or 'probacterial' effects in *in vitro* studies; however, *in vivo* evidence for such effects is so far lacking. The mechanisms by which autophagy genes mediate *in vivo* resistance to infection are not fully understood, but are likely to involve a combination of xenophagy, other autophagy-protein-dependent effects on microbial replication or survival, activation of innate and adaptive immune responses, and/or alterations in pathogen-induced cell death (Fig. 3).

An important question is whether this function of autophagy in broad resistance to infection with intracellular pathogens extends to humans. Recent human genetic studies provide some clues. The immunity-related GTPase human *IRGM*, which regulates autophagy-dependent clearance of mycobacteria *in vitro*[49], was identified as a genetic risk locus for tuberculosis in a West African population[50]. Numerous studies have shown a crucial role for autophagy in defence against mycobacterial infection in human cells[1], and a genome-wide analysis of host genes that regulate *Mycobacterium tuberculosis* replication demonstrated that a predominance of factors were autophagy regulators[51]. Thus, it is possible that autophagy has a central role in resistance to one of the most important global infectious diseases — tuberculosis. Mutations in *NOD2*, which encodes an intracellular pathogen-recognition receptor (nucleotide-binding oligomerization-domain-containing protein 2) that functions in bacterial autophagy[52,53], are also associated with susceptibility to infection with another mycobacterial agent, *Mycobacterium leprae*, the aetiological agent of leprosy[54]. An exciting future venture will be to confirm whether *IRGM*, *NOD2* and other autophagy-related genes are involved in resistance to infection with mycobacteria and other infections in further human populations and, if so, whether this resistance is mediated by autophagy.

Microbes fight back

Microbes undergo strong selective pressure to develop strategies to block host defence mechanisms; the number of such strategies is a surrogate measure of the importance of the host defence mechanism in immunity. As reviewed elsewhere[1,55], viruses and intracellular bacteria have evolved several ways to adapt to host autophagy. They can antagonize autophagy initiation or autophagosomal maturation, evade autophagic recognition, or use components of the autophagy pathway to facilitate their own replication or intracellular survival. An emerging theme is that microbial antagonism of autophagy not only blocks the xenophagic degradation of intracellular pathogens, but also blocks the functions of autophagy in innate and adaptive immunity. A relatively unexplored yet crucially important frontier is how microbial antagonism may contribute more broadly to the role of microbes in diseases characterized by defective autophagy, such as cancer, neurodegenerative diseases, ageing and, potentially, autoimmune and inflammatory diseases.

Viral strategies to shut off autophagy include the blockade of positive upstream regulators of autophagy (such as the IFN-inducible RNA-activated eIF2α protein kinase (PKR) signalling pathway), the activation of negative upstream regulators of autophagy (such as the nutrient-sensing TOR kinase signalling pathway) or direct antagonism of the autophagy machinery[55]. The overlapping functions of the eIF2α kinase signalling pathway in stress-induced general translational arrest,

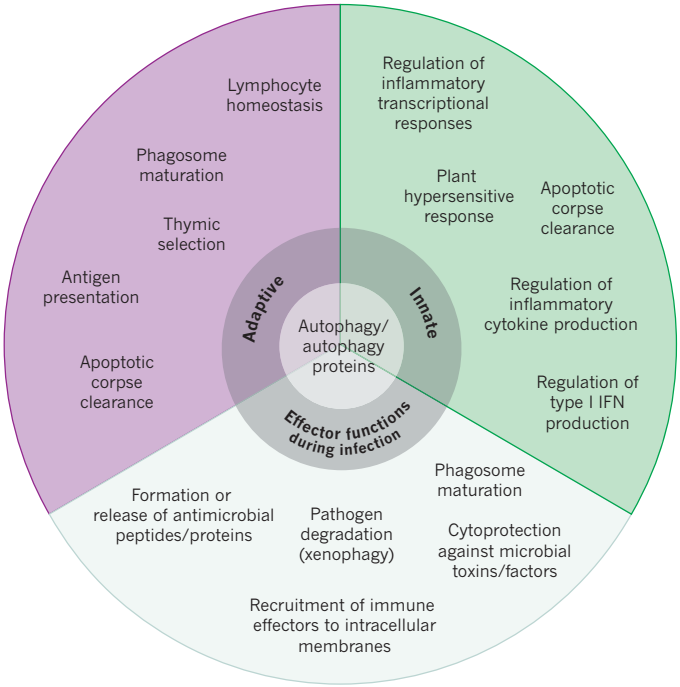


Figure 3 | Functions of the autophagy pathway and/or proteins in immunity. A summary of the known functions of the autophagy pathway and/or proteins in adaptive and innate immunity, and as effectors during infection.

transcriptional activation of stress-response genes and stress-induced autophagy enable viruses to disarm several facets of the cellular stress response to infection by one mechanism — that is, antagonism of eIF2α kinase signalling. The mTOR signalling pathway has a central role in the control of cell growth and metabolism, and interestingly, many of the viruses that activate mTOR are oncogenic (for example, Epstein–Barr virus, Kaposi's sarcoma-associated herpesvirus, hepatitis B virus and retroviruses). This suggests another type of pluripotent viral weapon — one that can promote oncogenesis by simultaneously inactivating autophagy and promoting cell growth through TOR activation. HIV envelope protein-dependent activation of mTOR signalling is also proposed to be a mechanism for HIV evasion of innate and adaptive immune responses in dendritic cells, including the degradation of incoming virions by lysosomes, blockade of HIV transfer to $CD4^+$ T cells, stimulation of TLR4 and TLR8 ligand responses, and presentation of HIV Gag antigen to $CD4^+$ T cells[56]. It will be important to determine whether these effects of HIV-mediated mTOR activation and autophagy inhibition contribute to impaired dendritic cell function during HIV infection *in vivo*.

Several viral proteins target the core autophagy protein beclin 1. Autophagosome initiation is blocked by interactions between beclin 1 and the herpes simplex virus 1 (HSV-1) neurovirulence factor ICP34.5 or the oncogenic γ-herpesvirus-encoded viral BCL2-like proteins, whereas autophagosome maturation is blocked by interactions between beclin 1 and the HIV accessory protein Nef or the influenza virus matrix protein 2 (ref. 55). The interactions between beclin 1, HSV-1 ICP34.5 and viral BCL2 are probably physiologically important *in vivo*; a mutant HSV-1 virus lacking the beclin-1-binding domain of ICP34.5 is attenuated in mouse models of encephalitis (presumably through its failure to control xenophagy and innate immunity)[57] and of corneal disease (through its failure to control adaptive immunity)[58]. Moreover, a mouse γ-herpesvirus that encodes a mutant viral BCL2 unable to bind to beclin 1 demonstrates impaired ability to maintain chronic infection[59]. Thus, viral antagonism of host autophagy can manipulate distinct aspects of viral pathogenesis and immunity *in vivo*.

It is not yet clear whether compared with other autophagy proteins,

beclin 1 is preferentially targeted by viral virulence proteins because of its central role in autophagosome formation, or more likely, whether we are just beginning to identify pairs of viral proteins and their autophagy pathway targets. In support of the latter, viral FLICE-like inhibitors encoded by Kaposi's sarcoma-associated herpesvirus and molluscum contagiosum virus suppress autophagy by interacting with the ATG3 E2-like enzyme, thereby preventing it from binding and processing LC3 (ref. 60).

Bacteria possess diverse strategies to avoid degradation by autophagolysosomal pathways. As reviewed elsewhere[1,38], many bacteria that reside in phagosomes or other vacuolar compartments have methods to inhibit lysosomal fusion or maturation, which, in the case of mycobacteria, can be partially overcome by treatments that stimulate autophagy. Another possible mechanism for bacterial evasion of autophagy has emerged from a genome-wide screen to identify host factors that regulate the intracellular survival of *M. tuberculosis*[51]. According to bioinformatics analyses, *M. tuberculosis* infection of human macrophage-like cells activates cellular pathways that inhibit autophagy. Intracellular bacteria that escape into the cytoplasm, such as *S. flexneri* and *L. monocytogenes*, use strategies to camouflage themselves to avoid autophagic recognition. The *Shigella* T3SS effector IcsB competitively binds to VirG, thereby preventing the interaction between ATG5 and VirG, a bacterial surface protein required for actin-based motility and *Shigella* targeting to autophagosomes[38]. The *Listeria* protein ActA interacts with cytosolic actin polymerization machinery (ARP2/3, VASP and actin), which blocks bacterial association with ubiquitin, p62 recruitment and autophagic targeting[61]. The precise mechanism of this block is unknown, but it has been proposed that the ActA-dependent recruitment of host cell cytoskeletal proteins may enable the bacterium to disguise itself as a host cell organelle[61]. This concept sheds light on the autophagy pathway in a fundamental aspect of immunology — the basis for discrimination between self and non-self.

Microbes have evolved not only to antagonize autophagy (as a cellular defence mechanism that threatens their survival), but also to exploit its components and functions in membrane trafficking for their own self-serving purposes[1,55]. An early concept in the field is that autophagosomes may serve as a protected niche for intracellular bacteria (provided fusion with acidic compartments is blocked) and/or serve as a source of nutrients for intracellular pathogens (which would require intact autophagolysosomal fusion)[1]. Trafficking of the intracellular bacteria *Yersinia pseudotuberculosis* to acidic compartments was recently shown to be enhanced by genetic inhibition of autophagy[62]. This seemingly contradicts other evidence that the autophagy proteins promote phagosomal maturation, but is consistent with the concept that autophagosomes function as a protected intracellular niche for bacteria. The role of the autophagy machinery in promoting and/or inhibiting vacuolar acidification — and the counter effects of microbes that reside in vacuolar compartments — needs to be explored further.

The function of autophagy proteins in membrane formation and/or trafficking is exploited by numerous viruses, including poliovirus, rotavirus, HIV, coronaviruses, Dengue virus, and the hepatitis B and C viruses[55]. Autophagosomes (defined as LC3-positive membranes, see caveat below) may act as a scaffold for intracellular membrane-associated replication of certain cytoplasmic RNA viruses[55]. Autophagy may assist in HIV biogenesis, because the processing of the HIV envelope precursor protein Gag and extracellular viral release are enhanced by the autophagy machinery[63]. Similarly, autophagy proteins are required for maximal levels of poliovirus egress[55]. Another newly defined proviral function of autophagy is its role in productive hepatitis C virus replication; several different autophagy proteins (such as beclin 1, ATG4B, ATG5, ATG7 and ATG12) assist in the translation of incoming, but not progeny, viral RNA[64]. ATG7 and class III PI(3)K activity also enhance hepatitis B virus DNA replication[65].

The mechanisms by which autophagy proteins facilitate the replication and/or egress of certain viruses are not yet understood. Some observations may relate to the role of autophagy proteins in remodelling the ER (vis-à-vis viral replication) or the role of autophagosomes in fusing with multivesicular bodies (vis-à-vis viral egress). It is possible that autophagy proteins function to provide membrane for viral replication complexes or translation machinery. This may be true for viruses such as hepatitis C virus, for which genetic knockdown of several different autophagy genes decreases productive replication[64]. However, for other viruses such as coronaviruses, the biogenesis of double-membrane, ER-derived vesicles used for replication proceeds through a pathway that involves the non-lipidated form of LC3 (LC3-I) but not the general autophagy machinery[66]. Thus, caution must be exercised in interpreting the significance of the co-localization (or biochemical interaction) of viral proteins and LC3, as LC3 may have autophagy-independent roles in membrane dynamics.

Autophagy regulation by immune signalling molecules

The central importance of autophagy in immunity is further underscored by the multitude of immune-related signalling molecules that regulate autophagy. As reviewed in detail elsewhere[2–4,38], autophagy is induced by different families of pathogen-recognition receptors (such as TLRs, NOD-like receptors and the double-stranded RNA-binding protein PKR), DAMPs (such as ATP, ROS and misfolded proteins), pathogen receptors (such as CD46), IFN-γ and downstream immunity-related GTPases, and DAP kinase, JNK, CD40, tumour necrosis factor-α (TNF-α), inhibitor of NF-κB (IKK) and NF-κB (Fig. 1). High mobility group box (HMGB) proteins have also been shown to function as both universal sensors of nucleic acids in innate immune signalling[67] and inducers of autophagy[68]. Autophagy is inhibited by BCL2, NF-κB, T helper 2 (T_H2) cytokines and the canonical nutrient-sensing insulin–AKT–TOR pathway. Inactivation of this nutrient-sensing pathway may contribute to vesicular stomatitis virus stimulation of autophagy in *Drosophila*[46], and autophagy activation in *C. elegans* with loss-of-function mutations in this pathway may mediate pathogen resistance in long-lived mutant nematodes[48]. Thus, both 'immune-specific' and more general nutrient-response signals control autophagy in response to infection.

Studies with vitamin D3 have uncovered a possible link between nutrition, innate immunity and the control of autophagy during mycobacterial infection. Low vitamin D3 levels are associated with increased susceptibility to tuberculosis. Vitamin D3 generates an antimycobacterial peptide, cathelicidin, and induces autophagy and mycobacterial killing in human monocytes through cathelicidin-dependent effects[69]. Although cathelicidin is required for vitamin-D3-dependent transcriptional upregulation of autophagy genes such as *BECN1* and *ATG5*, and vitamin D3 enhances the recruitment of cathelicidin to autophagosomes, it is not yet clear how cathelicidin promotes autophagy. Nonetheless, these observations may begin to provide some insight into the century-old Nobel prize award (Niels Ryberg Finsen, 1903) for the use of ultraviolet-light therapy (which generates active vitamin D3) in the treatment of diseases such as tuberculosis.

In most instances, the mechanisms of autophagy control by immune-related signalling molecules are not understood. However, there are some examples of specific interactions between immune signals and autophagy proteins that may be relevant to these mechanisms. For example, the interaction between beclin 1 and BCL2 (which inhibits its activity) is thought to be disrupted by the TLR adaptors MyD88 and TRIF, as well as by HMGB1, which bind to beclin 1 and displace BCL2 (refs 3, 68). Two intracellular sensors responsible for inducing autophagy in response to bacterial infection, NOD1 and NOD2, interact with ATG16L1 and recruit it to the plasma membrane, resulting in enhanced association of invasive bacteria (*S. flexneri*) with LC3 (ref. 53). Interestingly, a NOD2 mutation associated with Crohn's disease impairs ATG16L1 plasma membrane recruitment and bacterial co-localization with LC3 (ref. 53).

The identification of other possible protein–protein interactions between core autophagy proteins and immune signals by a large proteomics screen[21] has the potential to foster further advances in understanding how different immune signals regulate the autophagy machinery. For example, tectonin proteins with multivalent

β-propeller folds are known to function in pathogen recognition and innate immunity in invertebrates[70]. Thus, the interactions between previously uncharacterized human proteins of this tectonin family, TECPR1 and TECPR2, with the ATG5–ATG12–ATG16L1 complex and ATG8 family members, respectively[21], may contribute to pathogen-induced autophagy stimulation or selective autophagic targeting of pathogens in mammals.

Further links between immune signalling molecules and autophagy regulation were suggested by a genome-wide short interfering RNA screen[71]. The analysis identified 219 genes that suppressed basal autophagy, largely in a mammalian TOR complex 1 (mTORC1)-independent fashion. These included several cytokines such as CLCF1, LIF, IGF1, FGF2 and the chemokine SDF1 (also known as CXCL12), as well as cellular signalling molecules regulated by cytokines such as STAT3. These findings raise the possibility that cytokines may have a broader role in the control of autophagy than previously thought. Moreover, because these cytokine signalling pathways are important in immune cells, another central question is to what extent cytokine-mediated regulation of autophagy governs immune cell function. Given the general function of autophagy in cellular homeostasis[5], and the more specific functions in regulating immune and inflammatory signalling (discussed in 'Regulation of immune signalling by autophagy proteins'), cytokine-mediated changes in autophagy levels in immune cells may have a central role in immunity and inflammation.

Autophagy and adaptive immunity

Autophagy proteins function in adaptive immunity, including in the development and homeostasis of the immune system and in antigen presentation (Table 1 and Fig. 3). The knockout of different autophagy genes in specific lymphocyte populations in mice has shown a crucial role for autophagy proteins in the maintenance of normal numbers of B1a B cells, $CD4^+$ T cells, $CD8^+$ T cells and fetal haematopoietic stem cells[2,44,72]. In T cells, in which mitochondrial numbers are developmentally regulated during the transition from thymocyte to mature circulating T cell, the developmental defect in autophagy-deficient cells may be related to the defective clearance of mitochondria[44]. Another crucial function of autophagy in the development and homeostasis of the immune system is the elimination of autoreactive T cells in the thymus[44]. High levels of autophagy are present in thymic epithelial cells, in which autophagy participates in the delivery of self-antigens to MHC class II loading compartments. Genetic disruption of *Atg5* in thymic epithelial cells leads to the altered selection of certain MHC class II restricted T-cell specificities and autoimmunity[73]. Beyond these functions in lymphocyte survival and thymic negative selection, autophagy may exert other functions in lymphocyte differentiation, perhaps, in part, indirectly through effects on cytokine expression (see the next section). It is not yet known whether autophagy is involved in the development and/or homeostasis of immune cell populations other than lymphocytes and haematopoietic stem cells.

Autophagy proteins may participate in different facets of antigen presentation, including the delivery of endogenous antigens for MHC class II presentation to $CD4^+$ T cells[74,75], the enhancement of antigen donor cell cross-presentation to $CD8^+$ T cells[75], dendritic cell cross-presentation of phagocytosed antigens to $CD4^+$ T cells[32] and, in one report, MHC class I presentation of intracellular antigens to $CD8^+$ T cells[27]. The discovery that autophagosomes could deliver endogenous antigens to MHC class II loading compartments sheds light on one of the central mysteries of antigen presentation — how the immune system elicits $CD4^+$ T-cell responses to antigens that originate in all parts of the cell. The autophagic delivery of endogenously synthesized antigens for MHC class II presentation has been demonstrated *in vitro* for certain viral antigens[75], and probably explains the essential role of *Atg5 in vivo* in negative thymic selection[73]. However, the relative importance of this pathway in antigen presentation during infection *in vivo* is not yet known. There is nonetheless interest in exploiting this pathway for optimizing vaccine-elicited $CD4^+$ T-cell responses, by either pre-treating dendritic cells with autophagy-inducing agents in cell-based vaccine strategies or fusing antigens with LC3 to enhance their targeting to autophagosomes[1].

Of note, autophagy proteins are required for antigen cross-presentation during infection *in vivo*[32]. Dendritic-cell-specific deletion of *Atg5* in mice results in defects in priming $CD4^+$ T-cell responses after HSV and *Listeria* infections, and mice succumb more rapidly to lethal disease after intravaginal HSV infection. *Atg5*-deficient dendritic cells have normal migration, innate responses, endocytic and phagocytic activity and cross-presentation of peptides on MHC class I molecules. However, they exhibit defects in phagosome-to-lysosome fusion and in cross-presentation by MHC class II molecules of phagocytosed antigens containing TLR ligand. Thus, the interior of the phagosome, like that of the autophagosome, is a cellular compartment that autophagy-protein-dependent antigen presentation accesses to generate peptides for presentation to $CD4^+$ T cells. A potential evolutionary advantage of this autophagy-protein-dependent cross-presentation is that, by delegating antigen presentation duties to uninfected dendritic cells, the host can bypass the blockade of antigen presentation that may result from microbial antagonism of autophagy in infected cells.

Regulation of immune signalling by autophagy proteins

In response to infection, the host must activate those arms of the innate and adaptive immune system (including autophagy-dependent functions; Fig. 3) that help to control infection while, in parallel, triggering specific responses that limit detrimental, uncontrolled immune activation and inflammation. An exciting new frontier in autophagy research is the growing recognition of the function of autophagy proteins in achieving this balance (Fig. 4).

Autophagy proteins function in both the activation and inactivation of innate immune signalling[4]. The autophagy pathway activates type I IFN production in plasmacytoid dendritic cells by delivering viral nucleic acids to endosomal TLRs[76]. By contrast, autophagy proteins negatively regulate RIG-I-like receptor (RLR)-mediated induction of type I IFN production through the autophagic elimination of damaged mitochondria (and reduction of ROS)[77], and by the binding of ATG5–ATG12 to caspase recruitment domains of RLR signalling molecules[78]. Moreover, the autophagy protein ATG9A, but not ATG7, negatively regulates the activation of STING, a transmembrane protein that is required for efficient activation of type I IFN and pro-inflammatory cytokine production in response to stimulatory DNA[23]. Thus, it seems that autophagy proteins can negatively regulate IFN production by both autophagy-dependent and -independent mechanisms. With respect to the latter, different autophagy proteins may be specialized to target different innate immune signalling molecules.

The autophagy pathway and/or proteins also have a crucial role in the control of inflammatory signalling. A major effect is on the regulation of inflammatory transcriptional responses. Increased levels of the adaptor protein p62, which accumulates in autophagy-deficient cells, activate the pro-inflammatory transcription factor NF-κB through a mechanism involving TRAF6 oligomerization[79]. The accumulation of p62 in *Atg7*-deficient hepatocytes results in enhanced activity of the stress-responsive transcription factor NRF2 and NRF2-dependent liver injury[80]. In addition, Paneth cells (intestinal immune epithelial cells) from mice hypomorphic for *Atg16l1* (*Atg16l1*^{HM}) show enhanced transcription of pro-inflammatory cytokines and adipokines[81].

A second important effect of autophagy proteins on inflammatory signalling is at the level of the inflammasome. This complex contains NOD-like receptor cryopyrin proteins, the adaptor protein ASC and caspase 1, and is activated by cellular infection or other stress to promote the maturation of pro-inflammatory cytokines such as interleukin-1β (IL-1β) and IL-18 (ref. 4). *Atg16l1*- or *Atg7*-deficient mouse macrophages produce increased levels of mature IL-1β and IL-18 after TLR4 stimulation by endotoxin[82]. In addition, mouse chimaeras engrafted with *Atg16l1*$^{-/-}$ fetal liver haematopoietic progenitors have increased serum concentrations of IL-1β and IL-18

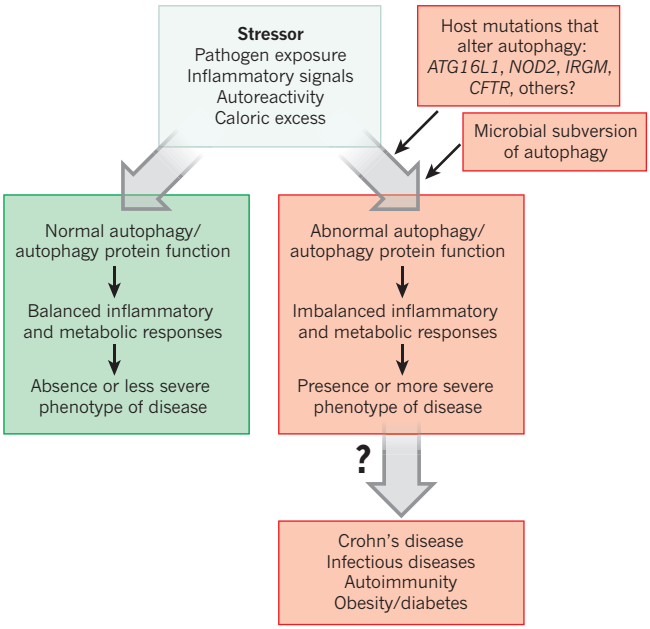


Figure 4 | Autophagy/autophagy proteins act to achieve a balance between activation and inactivation of innate immune signalling. A general model in which the levels of autophagy and autophagy proteins control disease in response to stressors. Normal autophagy protein function (green) contributes to balanced inflammatory and metabolic responses, resulting in protection against disease. Altered autophagy protein function (red) results in maladaptive inflammatory and metabolic responses, increased inflammation and more severe disease.

after treatment with dextran sodium sulphate (DSS), which contributes to increased DSS-induced colitis[82].

The mechanism(s) by which autophagy proteins negatively regulate inflammasome activation are not yet understood. Mutually non-exclusive possibilities include direct interactions between autophagy proteins and inflammasome components, indirect inhibition of inflammasome activity through autophagic suppression of ROS accumulation, or autophagic degradation of danger signals that activate the inflammasome. In line with the latter model, the autophagic degradation of amyotrophic-lateral-sclerosis-linked mutant superoxide dismutase has been proposed to limit caspase 1 activation and IL-1β production[83].

In addition to regulating inflammatory signalling, the autophagy pathway may prevent tissue inflammation through its role in apoptotic corpse clearance. The efficient clearance of apoptotic corpses during development and tissue homeostasis prevents secondary necrosis, which releases danger signals (DAMPs) that trigger inflammation. Autophagy genes are essential for the heterophagic clearance of dying apoptotic cells during developmental programmed cell death (by the generation of ATP-dependent engulfment signals)[84], and the retinas and lungs of embryonic mice lacking *Atg5* have a defect in apoptotic corpse engulfment that is associated with infiltration of inflammatory cells[84]. On the basis of growing evidence that autophagy proteins function in TLR-mediated phagolysosomal pathways, it is possible that autophagy also functions in phagocytes to facilitate apoptotic corpse clearance. Thus, in tissues such as the intestine, in which physiological regeneration involves continuous shedding or apoptosis of epithelial cells, autophagy-dependent functions in dying cells and/or phagocytic cells may promote efficient corpse clearance, thereby limiting inflammation.

Autophagy and inflammatory disease

Perturbations in autophagy-protein-dependent functions in immunity may contribute not only to increased susceptibility to infection, but also to chronic inflammatory diseases and autoimmune diseases. The only well-characterized link thus far is between mutations in autophagy regulators and Crohn's disease, a chronic inflammatory disorder of the small intestine, in which a breakdown in clearance or recognition of commensal bacteria, as well as altered mucosal barrier function and cytokine production, is thought to lead to intestinal inflammation (Fig. 5). Other emerging links include the autoimmune disease systemic lupus erythematosus (SLE), inflammation-associated metabolic diseases such as obesity and diabetes, and inflammation associated with cystic fibrosis lung disease (Fig. 4).

The role of autophagy proteins in Crohn's disease was not suspected until genome-wide association studies identified three Crohn's disease susceptibility genes, *IRGM*, *NOD2* and *ATG16L1*, that are involved in autophagy[85]. The *IRGM* risk allele contains a deletion in the promoter region of the gene that may be associated with changes in IRGM protein expression and may contribute to Crohn's disease, given IRGM's role in autophagy-dependent control of bacterial infection[49]. However, this hypothesis has not yet been tested. The three major Crohn's-disease-associated *NOD2* variants (a frameshift mutant and two missense mutants) may be loss-of-function mutants, with impaired muramyl dipeptide (MDP)-induced inflammatory signalling[86]. How the loss of function of a pro-inflammatory signal mechanistically contributes to an inflammatory disorder has been unclear, but the recently discovered links between NOD2 and autophagy may solve this conundrum. In primary immature human dendritic cells, NOD2 is required for MDP-induced autophagy, a process that is essential for the MHC class II presentation of bacterial antigens to $CD4^+$ T cells and for bacterial targeting to lysosomes[52]. Dendritic cells expressing Crohn's disease *NOD2* risk variants are defective in both of these functions[52]. Thus, in patients with Crohn's disease and *NOD2* risk variants, aberrant autophagy-dependent bacterial clearance and immune priming could act as a trigger for intestinal inflammation.

A mechanistic link may also exist between *ATG16L1* mutation and Crohn's disease pathogenesis. Similar to findings with *NOD2* variants, dendritic cells from patients with the Crohn's-disease-associated *ATG16L1*(*T300A*) risk variant are defective in presenting bacterial antigen to $CD4^+$ T cells[52]. However, it is not yet known how the T300A mutation affects the function of the mammalian ATG16L1 protein. This mutation resides in the carboxy-terminal WD-repeat domain that is absent in yeast Atg16 and is dispensable for autophagy. Although some studies have suggested that the *ATG16L1*(*T300A*) variant has reduced autophagic clearance of enteric pathogens such as adherent-invasive *Escherichia coli*[87] or *S.* Typhimurium[88], it remains controversial whether the risk versus protective alleles of *ATG16L1* have differences in stability or antibacterial autophagic activity[89].

Despite the uncertain nature of the effects of the T300A mutation on ATG16L1 function, *Atg16l1* mutation (null or hypomorphic alleles) in mice results in abnormalities that are relevant to Crohn's disease pathogenesis. As noted earlier, loss of Atg16l1 function in mice results in enhanced TLR-agonist-induced pro-inflammatory cytokine production by macrophages[82], enhanced DSS-induced colitis[82,90] and altered inflammatory gene transcriptional profiles in Paneth cells[81,90]. In addition, the Paneth cells of mice expressing low Atg16l1 levels (*Atg16l1*^{HM}) show defects in the packaging and extrusion of antimicrobial granules into the gut lumen; Paneth cells from patients with Crohn's disease and the *ATG16L1*(*T300A*) risk variant show similar defects[81]. This suggests that, in addition to the overlapping functions of NOD2 and ATG16L1 in a common bacterial-sensing pathway that promotes bacterial antigen presentation, ATG16L1 may have unique protective functions, including Paneth cell antimicrobial peptide release and the negative regulation of pro-inflammatory cytokine production. To connect the striking phenotypes in Atg16l1-mutant mice and the pathogenesis of Crohn's disease in humans with the *ATG16L1*(*T300A*) risk allele, the precise effects of the T300A mutation on ATG16L1 protein function need to be uncovered.

A new dimension in understanding the multifactorial basis of chronic inflammatory diseases such as Crohn's disease has emerged from the discovery that a virus trigger is required to observe intestinal

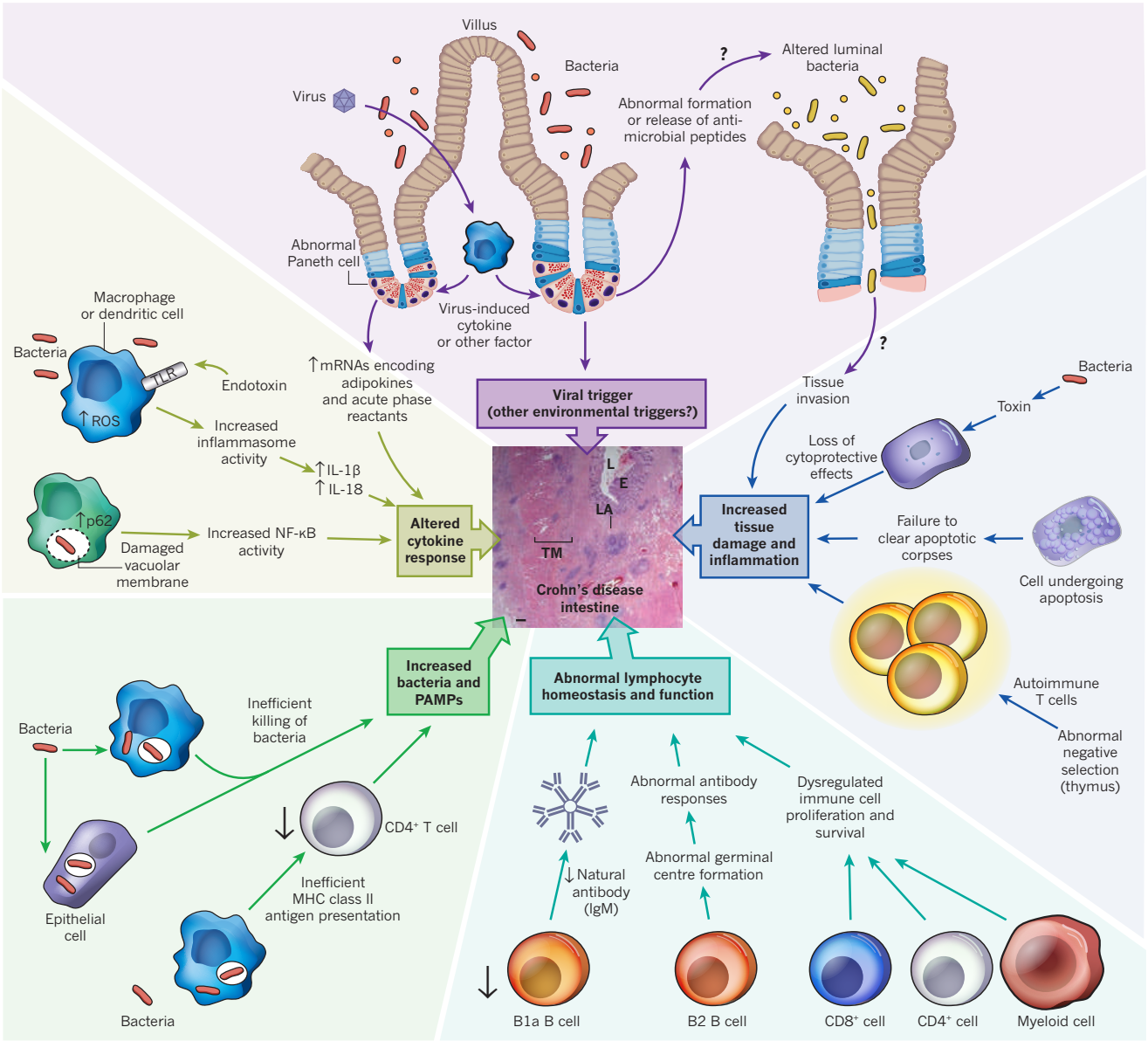


Figure 5 | The link between mutations in autophagy regulators and the chronic inflammatory disorder Crohn's disease. An overview of the many possible mechanisms by which defects in autophagy and autophagy protein function may contribute to the pathogenesis of a type of inflammatory bowel disease, Crohn's disease. A micrograph of a human small intestine from a patient with Crohn's disease is shown (centre), demonstrating the severe transmural inflammation that is characteristic of this disease. The postulated mechanisms by which defects in autophagy protein function might contribute to the development or perpetuation of intestinal inflammation are based on studies *in vitro* and animal models. There is no direct evidence that autophagy defects contribute to human Crohn's disease, although mutations in three autophagy-related genes, *ATG16L1*, *NOD2* and *IRGM*, are known to enhance risk of the disease. E, epithelium; IgM, immunoglobulin M; L, lumen; LA, lymphoid aggregates; TM, thickened muscle. Scale bar, 200 μm.

abnormalities in *Atg16l1*^{HM} mice[90]. In mice raised in a pathogen-free facility, only *Atg16l1*^{HM} mice (and not wild-type mice) infected with a virus found in routine conventional animal facilities, a murine norovirus, showed abnormal Paneth cell granule secretion, Paneth cell pro-inflammatory gene-expression profiles, and intestinal inflammation in response to DSS treatment[90]. This mucosal inflammation depended on the presence of the microbiome and pro-inflammatory cytokines, as it was reversed by antibiotic treatment or by TNF-α or IFN-γ inhibition. Thus, variations in a host autophagy gene, exposure to a specific virus and the microbiome can act together to trigger intestinal inflammation in mice that is similar to that in patients with Crohn's disease. Although environmental factors, including the gut microbiome, have long been suspected to contribute to Crohn's disease in genetically susceptible individuals, formal proof of this concept was lacking, and viruses were a previously unsuspected trigger. Another implication of this work is the concept that autophagy proteins, through their diverse roles in immunity and the control of inflammation, may serve as a central rheostat that prevents inflammatory diseases triggered by environmental stress (Fig. 4).

An important unanswered question is whether perturbations in autophagy may also result in inflammatory autoimmune disease. Genome-wide association studies have linked several single nucleotide polymorphisms (SNPs) in *ATG5* to SLE susceptibility[91–93]. SLE is a multifactorial, heterogeneous disease characterized by autoimmune responses against self-antigens generated from dying cells. Although the effects of these SNPs on ATG5 expression and function are not known, the lack of Atg5-dependent negative thymic selection generates autoimmunity and multi-organ inflammation in mice[73]. Loss of other ATG5-dependent effects, including regulation of IFN and pro-inflammatory cytokine secretion[77,78], clearance of dying cells[84] and dendritic cell

antigen presentation[32], might also contribute to the autoimmunity and inflammation associated with SLE. Thus, a link between *ATG5* mutation (or mutation of other autophagy genes) and SLE pathogenesis is biologically plausible, although not yet proven.

Defects in autophagy may contribute to inflammation-associated metabolic diseases such as diabetes and obesity, which are both linked to insulin resistance. The metabolic inflammasome — a complex composed of signalling molecules such as PKR, eIF2α, JNK, IRS and IKK— may act as a link between ER stress and more global stress responses, including inflammation and metabolic dysfunction (as observed in insulin resistance and obesity)[94]. Although most components of the metabolic inflammasome promote autophagy, the induction of autophagy by this signalling complex would be expected to serve as a negative-feedback mechanism that limits ER stress and disease progression. Consistent with this postulated protective effect of autophagy, hepatic suppression of the autophagy gene *Atg7* in mice results in increased ER stress and insulin resistance[95], and mice deficient in the autophagy adaptor protein p62 develop mature-onset obesity and insulin resistance[96]. Furthermore, obesity is associated with the accumulation and activation of macrophages and subsets of T cells in adipose tissue and the production of cytokines such as TNF-α and IL-6 (ref. 97). Thus, the failure of autophagy-dependent control of ER stress, immune cell homeostasis, immune cell activation and/or pro-inflammatory cytokine secretion may contribute to inflammation-associated responses that underlie the pathogenesis of metabolic diseases.

Another potential link between autophagy deficiency and chronic inflammation is in cystic fibrosis[98], a life-threatening genetic disorder caused by mutations in the gene encoding the cystic fibrosis transmembrane conductance regulator (CFTR). Mutations in *CFTR* lead to autophagy inhibition in lung epithelial cells through a mechanism that may involve ROS-mediated sequestration of the beclin 1–class III PI(3)K complex in perinuclear aggregates (redirecting it from its site of autophagy action at the ER). Restoration of beclin 1 and autophagy in cystic fibrosis epithelial cells rescues the disease phenotype, and antioxidants reverse the airway inflammation in a cystic fibrosis mouse model by a mechanism postulated to involve autophagy.

Future directions

The first series of studies demonstrating that the autophagy machinery is used to attack invading intracellular bacteria was published in 2004 (refs 30, 99, 100). Although autophagy had been observed at the ultrastructural level in cells infected with intracellular bacteria and viruses decades earlier, these studies were a seminal advance. For the first time, pharmacological and genetic manipulation of autophagy, which built on the discoveries of the yeast screens that identified the autophagy machinery, challenged the very notion of autophagy as an 'auto'- (self), 'phagy' (eating) pathway. Indeed, we learned that the same genes that are used to orchestrate the degradation of self-constituents, either for nutritional/energy homeostasis or cellular damage control, are also used to orchestrate the degradation of foreign invaders, termed xenophagy.

In the past few years, research in the field has uncovered new layers of complexity and functional diversity in terms of how this set of genes — originally characterized in the context of macroautophagy — may function to protect multicellular organisms against not only the threats of infection but also the threats of the host's own response to infection. The autophagy machinery does much more than form autophagosomes to engulf microbes — it somehow allows microbes in phagosomes and vacuoles to be targeted to the lysosome; it enables crucial cells in the immune system to develop properly and perform some of their 'normal' functions (such as produce IFN, secrete antimicrobial peptides or present antigens to stimulate adaptive immunity); and it ensures that these responses do not become out of control by functioning in central immunological tolerance and the negative regulation of innate and inflammatory signalling. Thus, recent advances may not only modify our understanding of immunity (in terms of understanding new roles of the autophagy machinery in immune regulation) but also reshape our understanding of the pathogenesis of inflammatory diseases (in terms of understanding how perturbations in autophagy protein function may contribute to such diseases).

Clearly, our understanding of the molecular mechanisms of the plethora of functions of autophagy proteins in immune-related processes is still quite primitive. We speculate that, similar to the way in which the initial genetic screens in yeast transformed autophagy research, current proteomic and genomic screens have the potential to transform research on autophagy and immunity. Such a transformation would include facilitating a much deeper understanding of the molecular mechanisms of the existing known immunological functions of autophagy through the use of the tools of modern systems biology to understand autophagy protein–protein interaction and signalling regulatory networks on a broad scale. Perhaps more exciting is the possibility that such a transformation will uncover new ways in which this ancient self-defence machinery can function in immunity. ■

1. Deretic, V. & Levine, B. Autophagy, immunity, and microbial adaptations. *Cell Host Microbe* **5,** 527–549 (2009).
2. Virgin, H. W. & Levine, B. Autophagy genes in immunity. *Nature Immunol.* **10,** 461–470 (2009).
3. Kroemer, G., Marino, G. & Levine, B. Autophagy and the integrated stress response. *Mol. Cell* **40,** 280–293 (2010).
4. Saitoh, T. & Akira, S. Regulation of innate immune responses by autophagy-related proteins. *J. Cell Biol.* **189,** 925–935 (2010).
5. Levine, B. & Kroemer, G. Autophagy in the pathogenesis of disease. *Cell* **132,** 27–42 (2008).
6. Mizushima, N., Yoshimori, T. & Levine, B. Methods in mammalian autophagy research. *Cell* **140,** 313–326 (2010).
This paper provides a concise and critical review of current methods to monitor and modulate autophagy in mammalian cells.
7. Schmid, D., Pypaert, M. & Munz, C. Antigen-loading compartments for major histocompatibility complex class II molecules continuously receive input from autophagosomes. *Immunity* **26,** 79–92 (2007).
8. Yu, L. *et al.* Termination of autophagy and reformation of lysosomes regulated by mTOR. *Nature* **465,** 942–946 (2010).
9. Nakatogawa, H., Suzuki, K., Kamada, Y. & Ohsumi, Y. Dynamics and diversity in autophagy mechanisms: lessons from yeast. *Nature Rev. Mol. Cell Biol.* **10,** 458–467 (2009).
10. Hayashi-Nishino, M. *et al.* A subdomain of the endoplasmic reticulum forms a cradle for autophagosome formation. *Nature Cell Biol.* **11,** 1433–1437 (2009).
11. Yla-Antilla, P., Vihinen, H., Jokitalo, E. & Eskelinen, E. L. 3D tomography reveals connections between the phagophore and endoplasmic reticulum. *Autophagy* **5,** 1180–1185 (2009).
12. Mizushima, N. The role of the Atg1/ULK1 complex in autophagy regulation. *Curr. Opin. Cell Biol.* **22,** 132–139 (2010).
13. Itakura, E. & Mizushima, N. Characterization of autophagosome formation site by a hierarchical analysis of mammalian Atg proteins. *Autophagy* **6,** 764–776 (2010).
14. Matsunaga, K. *et al.* Two Beclin 1-binding proteins, Atg14L and Rubicon, reciprocally regulate autophagy at different stages. *Nature Cell Biol.* **11,** 385–396 (2009).
15. Axe, E. L. *et al.* Autophagosome formation from membrane compartments enriched in phosphatidylinositol 3-phosphate and dynamically connected to the endoplasmic reticulum. *J. Cell Biol.* **182,** 685–701 (2008).
16. Polson, H. E. *et al.* Mammalian Atg18 (WIPI2) localizes to omegasome-anchored phagophores and positively regulates LC3 lipidation. *Autophagy* **6,** 506–522 (2010).
17. Ropolo, A. *et al.* The pancreatitis-induced vacuole membrane protein 1 triggers autophagy in mammalian cells. *J. Biol. Chem.* **282,** 124–133 (2007).
18. Tian, Y. *et al. C. elegans* screen identifies autophagy genes specific to multicellular organisms. *Cell* **141,** 1042–1055 (2010).
19. Fujita, N. *et al.* The Atg16L complex specifies the site of LC3 lipidation for membrane biogenesis in autophagy. *Mol. Biol. Cell* **19,** 2092–2100 (2008).
20. Weidberg, H. *et al.* LC3 and GATE-16/GABARAP subfamilies are both essential yet act differently in autophagosome biogenesis. *EMBO J.* **29,** 1792–1802 (2010).
21. Behrends, C., Sowa, M. E., Gygi, S. P. & Harper, J. W. Network organization of the human autophagy system. *Nature* **466,** 68–76 (2010).
22. Radoshevich, L. *et al.* ATG12 conjugation to ATG3 regulates mitochondrial homeostasis and cell death. *Cell* **142,** 590–600 (2010).
23. Saitoh, T. *et al.* Atg9a controls dsDNA-driven dynamic translocation of STING and the innate immune response. *Proc. Natl Acad. Sci. USA* **106,** 20842–20846 (2009).
24. Webber, J. L. & Tooze, S. A. New insights into the function of Atg9. *FEBS Lett.* **584,** 1319–1326 (2010).
25. Hailey, D. W. *et al.* Mitochondria supply membranes for autophagosome biogenesis during starvation. *Cell* **141,** 656–667 (2010).
26. Ravikumar, B., Moreau, K., Jahreiss, L., Puri, C. & Rubinsztein, D. C. Plasma membrane contributes to the formation of pre-autophagosomal structures. *Nature Cell Biol.* **12,** 747–757 (2010).
27. English, L. *et al.* Autophagy enhances the presentation of endogenous viral

antigens on MHC class I molecules during HSV-1 infection. *Nature Immunol.* **10,** 480–487 (2009).
28. Kraft, C., Peter, M. & Hofmann, K. Selective autophagy: ubiquitin-mediated recognition and beyond. *Nature Cell Biol.* **12,** 836–841 (2010).
29. Yamaguchi, H. *et al.* An initial step of GAS-containing autophagosome-like vacuoles formation requires Rab7. *PLoS Pathogens* **5,** e1000670 (2009).
30. Nakagawa, I. *et al.* Autophagy defends cells against invading group A *Streptococcus*. *Science* **306,** 1037–1040 (2004).
Reference 30, together with reference 99, provides the first evidence that autophagy has a key role in bacterial infection. Atg5 is shown to be essential for controlling the replication of group A *Streptococci* that escape into the cytoplasm.
31. Sanjuan, M. A. *et al.* Toll-like receptor signalling in macrophages links the autophagy pathway to phagocytosis. *Nature* **450,** 1253–1257 (2007).
32. Lee, H. K. *et al. In vivo* requirement for Atg5 in antigen presentation by dendritic cells. *Immunity* **32,** 227–239 (2010).
This study shows that the autophagy machinery is necessary for dendritic cells to process and present extracellular microbial antigens for MHC class II presentation *in vivo*, which protects mice against lethal viral infection.
33. Berger, S. B. *et al.* SLAM is a microbial sensor that regulates bacterial phagosome functions in macrophages. *Nature Immunol.* **11,** 920–927 (2010).
34. Huang, J. *et al.* Activation of antibacterial autophagy by NADPH oxidases. *Proc. Natl Acad. Sci. USA* **106,** 6226–6231 (2009).
35. Zhao, Z. *et al.* Autophagosome-independent essential function for the autophagy protein Atg5 in cellular immunity to intracellular pathogens. *Cell Host Microbe* **4,** 458–469 (2008).
36. Khaminets, A. *et al.* Coordinated loading of IRG resistance GTPases on to the *Toxoplasma gondii* parasitophorous vacuole. *Cell. Microbiol.* **12,** 939–961 (2010).
37. Zhao, Y. O., Khaminets, A., Hunn, J. P. & Howard, J. C. Disruption of the *Toxoplasma gondii* parasitophorous vacuole by IFNγ-inducible immunity-related GTPases (IRG proteins) triggers necrotic cell death. *PLoS Pathogens* **5,** 1–17 (2009).
38. Sumpter, R. Jr & Levine, B. Autophagy and innate immunity: triggering, targeting and tuning. *Semin. Cell Dev. Biol.* **21,** 699–711 (2010).
39. Shahnazari, S. *et al.* A diacylglycerol-dependent signaling pathway contributes to regulation of antibacterial autophagy. *Cell Host Microbe* **8,** 137–146 (2010).
40. Orvedahl, A. *et al.* Autophagy protects against Sindbis virus infection of the central nervous system. *Cell Host Microbe* **7,** 115–127 (2010).
41. Dupont, N. *et al. Shigella* phagocytic vacuolar membrane remnants participate in the cellular response to pathogen invasion and are regulated by autophagy. *Cell Host Microbe* **6,** 137–149 (2009).
42. Ponpuak, M. *et al.* Delivery of cytosolic components by autophagic adaptor protein p62 endows autophagosomes with unique antimicrobial properties. *Immunity* **32,** 329–341 (2010).
43. Guillemain, A. & Plus, N. Contrôle génique de la multiplication du virus de la sensibilité héréditaire au CO_2 chez *Drosophila melanogaster*. *Caryologia* (suppl.) 1211–1213 (1954).
44. Mizushima, N. & Levine, B. Autophagy in mammalian development and differentiation. *Nature Cell Biol.* **12,** 823–830 (2010).
45. Liu, Y. *et al.* Autophagy regulates programmed cell death during the plant innate immune response. *Cell* **121,** 567–577 (2005).
46. Shelly, S., Lukinova, N., Bambina, S., Berman, A. & Cherry, S. Autophagy is an essential component of *Drosophila* immunity against vesicular stomatitis virus. *Immunity* **30,** 588–598 (2009).
47. Yano, T. *et al.* Autophagic control of listeria through intracellular innate immune recognition in drosophila. *Nature Immunol.* **9,** 908–916 (2008).
48. Jia, K. *et al.* Autophagy genes protect against *Salmonella typhimurium* infection and mediate insulin signaling-regulated pathogen resistance. *Proc. Natl Acad. Sci. USA* **106,** 14564–14569 (2009).
49. Singh, S. B., Davis, A. S., Taylor, G. A. & Deretic, V. Human IRGM induces autophagy to eliminate intracellular mycobacteria. *Science* **313,** 1438–1441 (2006).
50. Intemann, C. D. *et al.* Autophagy gene variant *IRGM* –261T contributes to protection from tuberculosis caused by *Mycobacterium tuberculosis* but not by *M. africanum* strains. *PLoS Pathogens* **5,** e1000577 (2009).
51. Kumar, D. *et al.* Genome-wide analysis of the host intracellular network that regulates survival of *Mycobacterium tuberculosis*. *Cell* **140**, 731–743 (2010).
52. Cooney, R. *et al.* NOD2 stimulation induces autophagy in dendritic cells influencing bacterial handling and antigen presentation. *Nature Med.* **16,** 90–97 (2010).
53. Travassos, L. H. *et al.* Nod1 and Nod2 direct autophagy by recruiting ATG16L1 to the plasma membrane at the site of bacterial entry. *Nature Immunol.* **11,** 55–62 (2010).
54. Zhang, F. R. *et al.* Genomewide association study of leprosy. *N. Engl. J. Med.* **361,** 2609–2618 (2009).
55. Dreux, M. & Chisari, F. V. Viruses and the autophagy machinery. *Cell Cycle* **9,** 1295–1307 (2010).
56. Blanchet, F. P. *et al.* Human immunodeficiency virus-1 inhibition of immunoamphisomes in dendritic cells impairs early innate and adaptive immune responses. *Immunity* **32,** 654–669 (2010).
57. Orvedahl, A. *et al.* HSV-1 ICP34.5 confers neurovirulence by targeting the Beclin 1 autophagy protein. *Cell Host Microbe* **1,** 23–35 (2007).
58. Leib, D. A., Alexander, D. E., Cox, D., Yin, J. & Ferguson, T. A. Interaction of ICP34.5 with Beclin 1 modulates herpes simplex virus type 1 pathogenesis through control of $CD4^+$ T cell responses. *J. Virol.* **83,** 12164–12171 (2009).
59. E. X. *et al.* Viral Bcl-2-mediated evasion of autophagy aids chronic infection of γherpesvirus 68. *PLoS Pathogens* **5,** e1000609 (2009).
60. Lee, J. S. *et al.* FLIP-mediated autophagy regulation in cell death control. *Nature Cell Biol.* **11,** 1355–1362 (2009).
61. Yoshikawa, Y. *et al. Listeria monocytogenes* ActA-mediated escape from autophagic recognition. *Nature Cell Biol.* **11,** 1233–1240 (2009).
62. Moreau, K. *et al.* Autophagosomes can support *Yersinia pseudotuberculosis* replication in macrophages. *Cell. Microbiol.* **12,** 1108–1123 (2010).
63. Kyei, G. B. *et al.* Autophagy pathway intersects with HIV-1 biosynthesis and regulates viral yields in macrophages. *J. Cell Biol.* **186,** 255–268 (2009).
64. Dreux, M., Gastaminza, P., Wieland, S. F. & Chisari, F. V. The autophagy machinery is required to initiate hepatitis C virus replication. *Proc. Natl Acad. Sci. USA* **106,** 14046–14051 (2009).
This study shows that many autophagy proteins are essential for the initial stages of hepatitis C viral RNA translation, illustrating that autophagy proteins may be subverted to enhance the replication of intracellular pathogens.
65. Sir, D. *et al.* The early autophagic pathway is activated by hepatitis B virus and required for viral DNA replication. *Proc. Natl Acad. Sci. USA* **107,** 4383–4388 (2010).
66. Reggiori, F. *et al.* Coronaviruses hijack the LC3-I-positive EDEMosomes, ER-derived vesicles exporting short-lived ERAD regulators, for replication. *Cell Host Microbe* **7,** 500–508 (2010).
67. Yanai, H. *et al.* HMGB proteins function as universal sentinels for nucleic-acid-mediated innate immune responses. *Nature* **462,** 99–103 (2009).
68. Tang, D. *et al.* Endogenous HMGB1 regulates autophagy. *J. Cell Biol.* **190,** 881–892 (2010).
69. Yuk, J. M. *et al.* Vitamin D3 induces autophagy in human monocytes/macrophages via cathelicidin. *Cell Host Microbe* **6,** 231–243 (2009).
70. Low, D. H. *et al.* A novel human tectonin protein with multivalent β-propeller folds interacts with ficolin and binds bacterial LPS. *PLoS ONE* **4,** e6260 (2009).
71. Lipinski, M. M. *et al.* A genome-wide siRNA screen reveals multiple mTORC1 independent signaling pathways regulating autophagy under normal nutritional conditions. *Dev. Cell* **18,** 1041–1052 (2010).
72. Liu, F. *et al.* FIP200 is required for the cell-autonomous maintenance of fetal hematopoietic stem cells. *Blood* **116,** 4806–4814 (2010).
73. Nedjic, J., Aichinger, M., Emmerich, J., Mizushima, N. & Klein, L. Autophagy in thymic epithelium shapes the T-cell repertoire and is essential for tolerance. *Nature* **455,** 396–400 (2008).
This study provided the first evidence that the autophagy machinery functions in MHC class II antigen presentation *in vivo*, specifically in shaping the $CD4^+$ T-cell repertoire during negative thymic selection, and thereby preventing autoimmunity and multi-organ inflammation.
74. Paludan, C. *et al.* Endogenous MHC class II processing of a viral nuclear antigen after autophagy. *Science* **307,** 593–596 (2005).
This study provided the first evidence that the autophagy machinery can deliver endogenously synthesized antigens for presentation on MHC class II molecules to $CD4^+$ T cells.
75. Munz, C. Antigen processing via autophagy—not only for MHC class II presentation anymore? *Curr. Opin. Immunol.* **22,** 89–93 (2010).
76. Lee, H. K., Lund, J. M., Ramanathan, B., Mizushima, N. & Iwasaki, A. Autophagy-dependent viral recognition by plasmacytoid dendritic cells. *Science* **315,** 1398–1401 (2007).
77. Tal, M. C. *et al.* Absence of autophagy results in reactive oxygen species-dependent amplification of RLR signaling. *Proc. Natl Acad. Sci. USA* **106,** 2770–2775 (2009).
78. Jounai, N. *et al.* The Atg5–Atg12 conjugate associates with innate antiviral immune responses. *Proc. Natl Acad. Sci. USA* **104,** 14050–14055 (2007).
79. Moscat, J. & Diaz-Meco, M. T. p62 at the crossroads of autophagy, apoptosis, and cancer. *Cell* **137,** 1001–1004 (2009).
80. Komatsu, M. *et al.* The selective autophagy substrate p62 activates the stress responsive transcription factor Nrf2 through inactivation of Keap1. *Nature Cell Biol.* **12,** 213–223 (2010).
81. Cadwell, K. *et al.* A key role for autophagy and the autophagy gene *Atg16l1* in mouse and human intestinal Paneth cells. *Nature* **456,** 259–263 (2008).
82. Saitoh, T. *et al.* Loss of the autophagy protein Atg16L1 enhances endotoxin-induced IL-1β production. *Nature* **456,** 264–268 (2008).
This study demonstrates that autophagy proteins negatively control endotoxin-induced inflammasome activation. References 81 and 82 also show that autophagy gene deficiency increases susceptibility to experimentally induced inflammatory bowel disease.
83. Meissner, F., Molawi, K. & Zychlinsky, A. Mutant superoxide dismutase 1-induced IL-1β accelerates ALS pathogenesis. *Proc. Natl Acad. Sci. USA* **107,** 13046–13050 (2010).
84. Qu, X. *et al.* Autophagy gene-dependent clearance of apoptotic cells during embryonic development. *Cell* **128,** 931–946 (2007).
This paper shows that autophagy genes are necessary for apoptotic cells to generate engulfment signals required for successful apoptotic corpse clearance and the prevention of tissue inflammation.
85. Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genet.* **40,** 955–962 (2008).
86. Brain, O., Allan, P. & Simmons, A. NOD2-mediated autophagy and Crohn disease. *Autophagy* **6,** 412–414 (2010).
87. Lapaquette, P., Glasser, A. L., Huett, A., Xavier, R. J. & Darfeuille-Michaud, A. Crohn's disease-associated adherent-invasive *E. coli* are selectively favoured by impaired autophagy to replicate intracellularly. *Cell. Microbiol.* **12,** 99–113 (2010).

88. Kuballa, P., Huett, A., Rioux, J. D., Daly, M. J. & Xavier, R. J. Impaired autophagy of an intracellular pathogen induced by a Crohn's disease associated ATG16L1 variant. *PLoS ONE* **3,** e3391 (2008).
89. Fujita, N. *et al.* Differential involvement of Atg16L1 in Crohn disease and canonical autophagy: analysis of the organization of the Atg16L1 complex in fibroblasts. *J. Biol. Chem.* **284,** 32602–32609 (2009).
90. Cadwell, K. *et al.* Virus-plus-susceptibility gene interaction determines Crohn's disease gene *Atg16L1* phenotypes in intestine. *Cell* **141,** 1135–1145 (2010).
This study shows that both hypomorphic expression of an autophagy protein and a viral-infection trigger are necessary for experimentally induced inflammatory bowel disease, suggesting that the interaction between host defects in autophagy and environmental stressors such as infection may be crucial for the pathogenesis of certain inflammatory diseases.
91. Harley, J. B. *et al.* Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in *ITGAM*, *PXK*, *KIAA1542* and other loci. *Nature Genet.* **40,** 204–210 (2008).
92. Gateva, V. *et al.* A large-scale replication study identifies *TNIP1*, *PRDM1*, *JAZF1*, *UHRF1BP1* and *IL10* as risk loci for systemic lupus erythematosus. *Nature Genet.* **41,** 1228–1233 (2009).
93. Han, J. W. *et al.* Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nature Genet.* **41,** 1234–1237 (2009).
94. Nakamura, T. *et al.* Double-stranded RNA-dependent protein kinase links pathogen sensing with stress and metabolic homeostasis. *Cell* **140,** 338–348 (2010).
95. Yang, L., Li, P., Fu, S., Calay, E. S. & Hotamisligil, G. S. Defective hepatic autophagy in obesity promotes ER stress and causes insulin resistance. *Cell Metab.* **11,** 467–478 (2010).
96. Rodriguez, A. *et al.* Mature-onset obesity and insulin resistance in mice deficient in the signaling adapter p62. *Cell Metab.* **3,** 211–222 (2006).
97. Hotamisligil, G. S. Inflammation and metabolic disorders. *Nature* **444,** 860–867 (2006).
98. Luciani, A. *et al.* Defective CFTR induces aggresome formation and lung inflammation in cystic fibrosis through ROS-mediated autophagy inhibition. *Nature Cell Biol.* **12,** 863–875 (2010).
99. Gutierrez, M. G. *et al.* Autophagy is a defense mechanism inhibiting BCG and *Mycobacterium tuberculosis* survival in infected macrophages. *Cell* **119,** 753–766 (2004).
Reference 99, together with reference 30, provides the first evidence that autophagy has a key role in bacterial infection. Autophagy induction by IFN-γ, rapamycin or starvation results in the conversion of mycobacterial phagosomes into phagolysosomes, thereby enhancing mycobacterial killing.
100. Ogawa, M. *et al.* Escape of intracellular *Shigella* from autophagy. *Science* **307,** 727–731 (2005).

Acknowledgements The work in the authors' laboratories was supported by National Institutes of Health (NIH) grants RO1 CA109618 and U54 AI057156 (B.L.); by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology, Japan, and by the Takeda Science Foundation (N.M.); and by NIH grants RO1 AI054483, U54 AI057160, RO1 AI084887 and RO1 CA096511 and the Broad Medical Foundation (H.W.V.). We thank T. Stappenbeck for discussions, and A. Diehl and M. Harstein for scientific illustration. We apologize to those authors whose work could not be cited owing to space limitations.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing interests. Correspondence should be addressed to B.L. (beth.levine@utsouthwestern.edu) or H.W.V. (virgin@wustl.edu).

REVIEW

doi:10.1038/nature09783

Pervasive roles of microRNAs in cardiovascular biology

Eric M. Small[1] & Eric N. Olson[1]

First recognized as regulators of development in worms and fruitflies, microRNAs are emerging as pivotal modulators of mammalian cardiovascular development and disease. Individual microRNAs modulate the expression of collections of messenger RNA targets that often have related functions, thereby governing complex biological processes. The wide-ranging functions of microRNAs in the cardiovascular system have provided new perspectives on disease mechanisms and have revealed intriguing therapeutic targets, as well as diagnostics, for a variety of cardiovascular disorders.

Diseases of the cardiovascular system are the most common congenital birth defects and causes of adult morbidity and mortality[1–3]. Although the cellular mechanisms and gene mutations responsible for numerous cardiovascular disorders have been extensively studied, it has become apparent only recently that microRNAs (miRNAs) have key roles in cardiovascular development and disease[4–8]. The prominent functions of miRNAs in cardiovascular biology probably reflect the sensitivity of the cardiovascular system to relatively subtle perturbations in gene expression, which can result in severe and often fatal abnormalities.

A primary role of miRNAs seems to be the 'fine-tuning' of gene expression to control development and tissue homeostasis[8]. However, under conditions of stress, the functions of miRNAs become especially pronounced, underscoring their roles in disease. Highly specific patterns of miRNA expression correlate with different cardiovascular disorders (such as cardiac hypertrophy, heart failure[9–12], post-myocardial infarction remodelling[13,14] and vascular remodelling[15,16]), and gain- and loss-of-function miRNA studies in mice have revealed pathogenic and protective functions of miRNAs *in vivo*[6,17]. Correlation of the cellular targets of miRNA action with cardiovascular phenotypes illuminates new biological pathways and disease mechanisms. Especially intriguing is the ability to manipulate individual miRNAs *in vivo* using oligonucleotide-based inhibitors or miRNA mimics, thereby opening up possibilities for the therapeutic manipulation of miRNAs[18].

In this Review, we describe the biology and mechanisms of action of miRNAs in the cardiovascular system, and consider the opportunities and challenges for the therapeutic modulation of miRNAs in cardiovascular disease. See refs 8, 19 and 20 for more detailed reviews of the biosynthesis and mechanisms of action of miRNAs.

Functional concepts of miRNA action

MicroRNAs are ~22-nucleotide single-stranded RNAs that inhibit the expression of specific mRNA targets through Watson–Crick base pairing between the miRNA 'seed region' and sequences commonly located in the 3′ untranslated regions (UTRs). The human genome is estimated to encode up to 1,000 miRNAs[21], which are either transcribed as stand-alone transcripts, frequently encoding several miRNAs, or generated by the processing of introns of protein-coding genes[21]. The integration of miRNAs into introns of protein-coding genes serves to coordinate the expression of the miRNA with the mRNA encoded by that gene, without the necessity for a separate set of *cis*-regulatory elements to drive expression of the miRNA (Fig. 1a). It is not uncommon for intronic miRNAs to modulate the same biological processes as the protein encoded by the host gene[22–26]. The dual functions of such genes, encoding protein and miRNA, provide sophisticated feedback and feedforward regulatory networks, specific examples of which are highlighted throughout this Review.

Genetic deletions of miRNAs in organisms ranging from worms to mice have shown that few developmental processes are absolutely dependent on single miRNAs[8,27]. A recent study using compound mutant worms suggested there was significant redundancy within miRNA families, between unrelated miRNAs, and even between miRNAs and transcription factors, perhaps evolving as a buffer against deleterious variations in gene-expression programs[28,29]. The actions of miRNAs often become pronounced under conditions of physiological or pathological signalling, suggesting conditional activities of miRNAs that necessitate genetic perturbation or sensitizing agents to uncover their functions.

miRNAs typically exert modest inhibitory effects on many mRNAs, which often encode proteins that govern the same biological process — for example, the fibrotic response is inhibited by miR-29 (ref. 14), cardiac conduction by miR-1 (refs 30–32), actin cytoskeletal dynamics by miR-145 (ref. 16), the phosphatidylinositol-3-OH kinase (PI(3)K)–AKT pathway by miR-486 (ref. 33) and stem-cell pluripotency by miR-145 (ref. 34). The cumulative reduction in expression of several components of a molecular pathway reduces the importance of a single miRNA–mRNA interaction to elicit a biological response, and adds robustness to gene-regulatory networks (Fig. 1b). The multiplicity of miRNA targets may also promote combinatorial regulation by miRNAs that individually target various mRNAs whose protein products contribute to one particular regulatory axis (Fig. 1c). In this model, a biological response would be expected only after co-expression of several miRNAs that cooperatively target various components of a functional network or are all required to sufficiently repress a single target. By contrast, some miRNAs seem to reinforce an appropriately 'balanced' pathway by targeting both positive and negative regulatory components (for example, agonism and antagonism of Nodal signalling by miR-430)[35] (Fig. 1d). This mode of action allows buffering against minor physiological variations. Clearly, miRNA biology is a complex and highly orchestrated mode of gene regulation, potentially impinging on nearly all biological processes in mammals and having particularly important roles in disease states.

Oligonucleotide modulation of miRNA function

The ability of miRNAs to modulate important biological pathways offers opportunities for the manipulation of miRNA function using oligonucleotide inhibitors (antimiRs) or miRNA mimics (Fig. 2). Antisense

[1]Department of Molecular Biology, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9148, USA.

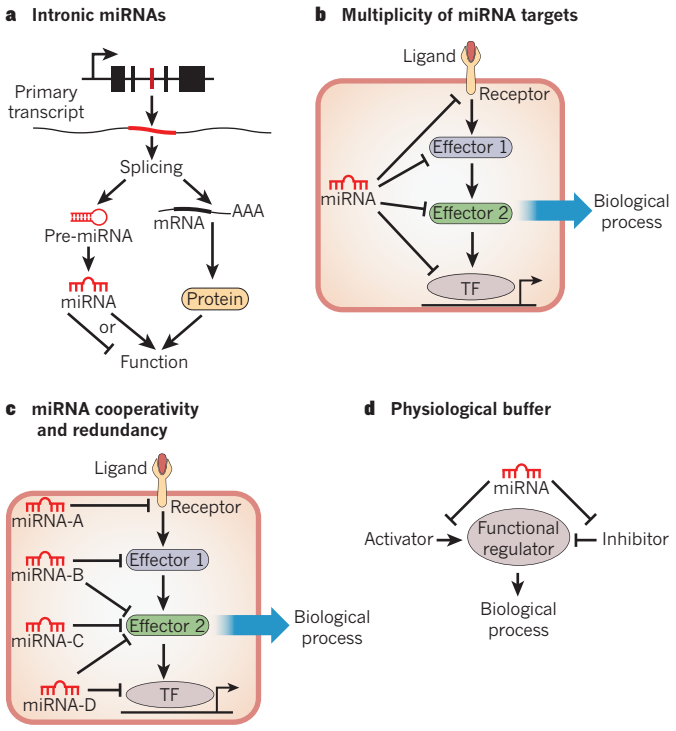


Figure 1 | Concepts of miRNA function. The potential modes of miRNA-based regulation of gene expression are shown. **a**, Intronic miRNAs are encoded within an intron of a host gene. mRNA splicing generates a protein-coding transcript and an miRNA stem–loop. Intronic miRNAs often regulate similar processes to that of the protein encoded by the host gene. AAA, polyadenylated tail of the transcript; pre-miRNA, precursor miRNA. **b**, A common mechanism of miRNA function involves the modest repression of several mRNAs in a common biological process by a single miRNA. This mechanism reduces the dependence on a single miRNA–mRNA interaction and increases the robustness of the gene-regulatory network. TF, transcription factor. **c**, Many miRNAs may cooperatively or redundantly regulate a single biological process, by individually targeting many components of that process or by synergistically repressing a crucial component of a pathway. **d**, miRNAs may act as a 'buffer' against minor perturbations in a biological pathway. This is accomplished by the targeting of factors that positively and negatively influence a particular process, thereby insulating that process from environmental fluctuations.

oligonucleotides directed against specific miRNA sequences are efficiently taken up by a variety of tissues and block miRNA function in the heart and vasculature[18].

Other oligonucleotide-based techniques involve 'target protectors' or 'masks', which block individual miRNAs from binding to their mRNA targets, thereby rescuing the mRNA from inhibition. Target protectors have been validated in zebrafish[35] and in cultured cardiac myocytes[36]. miRNA 'sponges' or 'decoys' containing several miRNA-binding sites also act as competitive inhibitors for miRNA binding[37,38]. Although the results obtained from pharmacological knockdown or overexpression of miRNAs sometimes differ from those obtained using genetic mouse models, the development of these various technologies has greatly accelerated the rate at which basic biological questions can be answered in an experimental setting.

miRNAs in cardiovascular development

The requirement of miRNAs for cardiovascular development and function was initially demonstrated by tissue-specific deletion in mice of the *Dicer* gene, which encodes an enzyme that is essential for miRNA processing. Lethal phenotypes were observed after *Dicer* deletion in myocardial and vascular lineages[39,40]. Although these findings highlight the crucial roles of miRNAs in the cardiovascular system, no specific miRNA deletion has yet been found to cause fully penetrant embryonic lethality in mice, indicating significant redundancy of miRNA function[28] and suggesting that the lethal consequences of *Dicer* deletion reflect the collective functions of many miRNAs rather than any single miRNA. The rapidly expanding number of miRNAs implicated in various aspects of cardiovascular biology precludes an in-depth review of all of them, so general principles of miRNA regulation and function are considered throughout this Review.

Roles of miRNAs in heart development

Heart formation requires precise and complex interactions among diverse cell types from several lineages — cardiomyocytes, endocardial, epicardial and vascular cells, fibroblasts and cells of the conduction system. Specific miRNAs are enriched in different cardiac cell types and, in some cases, have been found to participate in the specification of cell identity. Genetic ablation and antisense oligonucleotide-mediated knockdown studies have shown miRNA contributions to developmental processes as diverse as embryonic stem (ES)-cell differentiation, cardiomyocyte proliferation, contractility, ion-channel regulation and cardiac conduction (Fig. 3).

Expression profiling has shown that the 18 most abundant miRNAs in the heart account for more than 90% of all cardiac miRNAs[41]. Because a threshold level of miRNA expression seems to be required for the efficient repression of target gene expression (typically >100 copies per cell)[37,42], the regulation of heart development may depend on either a relatively discrete set of miRNAs or the combinatorial function of a larger array of miRNAs expressed at a low level. So far, a functional role in heart development has been ascribed to only the most enriched miRNAs, reflecting a dosage requirement, functional redundancy or both.

miR-1 is the most abundant miRNA in cardiac myocytes, and it was the first miRNA implicated in heart development[30]. miR-1 and the related miRNA miR-133 arise from a common precursor RNA, the expression of which in the embryonic heart is mediated by two separate enhancers that are regulated by the transcription factors SRF and MEF2 (refs 43, 44), integrating these miRNAs into well-characterized transcriptional networks. miR-1 and miR-133 seem to function cooperatively to promote mesoderm differentiation of ES cells and suppress endodermal and ectodermal cell fates[45]. By contrast, they have opposing roles later in the cardiac lineage when miR-1 promotes and miR-133 inhibits cardiomyocyte differentiation[45]. Neither miR-1 nor miR-133 is absolutely required for the specification of cardiac cell fates *in vivo*, as 50% of mice lacking either miRNA are viable[30,46]. This disparity between the functions of miRNAs as determined by *in vitro* assays versus *in vivo* loss-of-function studies is a common theme, and suggests that compensatory mechanisms that account for the unexpectedly mild phenotypes may be activated in genetic knockout mice. Zebrafish seem to be particularly sensitive to miRNA regulation, such that inhibition of specific miRNAs evokes more dramatic phenotypes than those seen in mutant mice. For example, antisense-mediated knockdown of miRNAs in zebrafish has revealed roles for miRNAs in the formation of the cardiac chambers and the atrioventricular canal[47,48].

Roles of miRNAs in vascular and blood development

The formation and function of the vascular system requires the establishment and remodelling of a contiguous series of lumenized tubes made of endothelial cells. Concurrent recruitment of vascular smooth muscle cells (SMCs) to the endothelial plexus during vessel maturation imparts the necessary tone and contractility for proper blood flow. Numerous miRNAs have been shown to govern these processes during vascular development and disease (Fig. 4).

The endothelial-cell-specific miRNA miR-126 is encoded by an intron of the epidermal growth factor-like domain 7 (*Egfl7*) gene, which encodes an endothelial-cell-enriched growth factor involved in the control of cell migration[49]. miR-126 is induced by blood flow and controls angiogenic sprouting of aortic arch vessels by the stimulation

of vascular endothelial growth factor signalling[50]. Mice lacking miR-126 are partially viable but have fragile and leaky blood vessels and defects in angiogenesis[51,52]. Antisense-oligonucleotide-mediated knockdown of miR-126 in zebrafish causes complete embryonic lethality owing to the loss of vascular integrity and haemorrhaging[53]. Vascular patterning in mouse retinas is also modulated by miR-218, which is encoded by an intron of the *Slit1* and *Slit2* genes and inhibits several components of the SLIT–ROBO signalling pathway[54]. This study is an important demonstration of the coordinated regulation of a biological process by an miRNA and its host gene.

miR-143 and miR-145, encoded by a bicistronic pre-miRNA, are expressed specifically in SMCs under the control of SRF and members of the myocardin family of co-activators. These miRNAs target numerous regulators of actin signalling, including Rho GTPases, slingshot homologue 2, adducin, cofilin and actin itself[16]. miR-145 has been reported to be necessary and sufficient for SMC differentiation *in vitro*[55]. However, mice lacking both miR-143 and miR-145 are viable, suggesting that further mechanisms modulate their functions *in vivo*[16,56,57]. miR-145-mutant mice have reduced vascular tone, which contributes to a reduction in blood pressure[16,57]. Vascular SMCs from these mutant mice show diminished sensitivity to mechanical injury and an abnormality in phenotypic switching in response to injury that seems to reflect perturbations in actin signalling and SRF activation. Collectively, these studies demonstrate that miRNAs function as sensors of mechanical and environmental changes, thus linking dynamic physiological processes with the regulation of gene expression.

The differentiation of blood cells is also dependent on miRNA activity[58]. An miRNA expression signature has been described for haematopoietic stem-cell progenitors, which show dynamic regulation during differentiation[59]. Furthermore, the *Ago2*- (also known as *Eif2c2*-) null mouse has erythroid lineage defects[60], and modulation of miRNA expression in erythroid progenitors suggests a role for miRNAs in their differentiation[61,62]. Indeed, loss-of-function studies in mice have also implicated miRNAs, including miR-223 and miR-451, in erythroid proliferation and differentiation[63–66]. For example, gene targeting and pharmacological knockdown of miR-451, which is enriched in erythroid cells, results in reduced baseline haematocrit levels and impaired erythroid expansion in response to oxidative stress[63,65,66]. Further analysis of miRNAs in circulating cells may reveal roles in functions such as oxygen delivery, angiogenesis and the inflammatory process.

miRNAs in cardiovascular disease

Heart failure and several cardiovascular diseases are associated with the re-expression of the fetal cardiac gene program, which may have causative or adaptive roles[1] and includes a signature pattern of miRNAs[10,12]. Indeed, numerous cardiac-enriched miRNAs show dynamic regulation in human heart disease, suggesting their involvement in the regulation of cardiovascular disease[9,11,12].

Roles of miRNAs in heart disease

The importance of individual miRNAs in the setting of heart disease has been shown by genetic deletion in mice subjected to various cardiovascular insults (Fig. 3). miRNAs are implicated in pathologies as diverse as arrhythmias (miR-1 (ref. 31), miR-133 (ref. 32) and miR-208a (ref. 67)), fibrosis (miR-21 (ref. 68) and miR-29 (ref. 14)), pressure-overload-induced remodelling (miR-208 (refs 67, 69) and miR-133 (ref. 70)), and metabolic disorders (miR-33 (ref. 24)).

One of the best-characterized examples of stress-dependent gene regulation by an miRNA involves a family of miRNAs encoded by myosin heavy chain (*MHC*) genes, referred to as MyomiRs[67,69,71]. This is also an example of intronic miRNAs participating in a process related to host gene function. Three members of this miRNA family, miR-208a, miR-208b and miR-499, are encoded by the α-*MHC* (also known as *Myh6*), β-*MHC* (*Myh7*) and *Myh7b* genes, respectively. These MyomiRs regulate a collection of transcriptional repressors and signalling molecules that govern MHC expression, as well as thyroid hormone activity and the stress-responsiveness of cardiac muscle cells. Deletion of *Mir208a* in mice abrogates the re-activation of the fetal β-*MHC* gene in response to haemodynamic cardiac stress, and protects the heart from pathological remodelling[67,69]. The MyomiR family thus constitutes an intricate regulatory circuit that controls myosin gene expression and cardiac stress responsiveness during adaptation to pathological signalling.

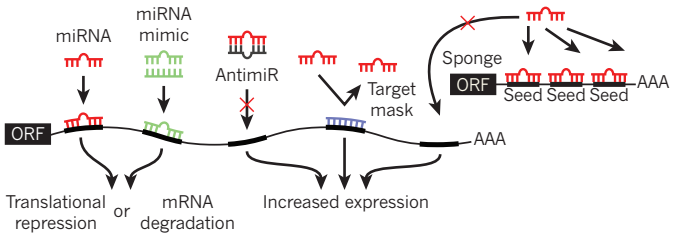


Figure 2 | Oligonucleotide manipulation of miRNA function. The various methods of artificially modulating miRNA expression or activity are shown. Endogenous miRNA (red) binds to complementary sequences in the 3′ UTR of a target gene, resulting in translational repression or mRNA degradation. An miRNA mimic (green) consists of an oligonucleotide duplex of the miRNA and a passenger strand. The miRNA mimic comprises the same nucleotide sequence as an endogenous miRNA, and is designed to target the same mRNAs as that miRNA. An antimiR (grey) is an oligonucleotide that is complementary to an endogenous miRNA, thereby designed to bind and inhibit its function. A target mask (blue) is an oligonucleotide designed to bind to a portion of an endogenous miRNA target without initiating mRNA degradation or translational inhibition. This strategy rescues one particular mRNA from miRNA-mediated repression. miRNA sponges consist of an open reading frame (ORF) linked to a 3′ UTR that contains several binding sites for a particular miRNA, acting as competitive inhibitors for miRNA binding.

In addition to the miR-208 family, several other miRNAs have been implicated as either causative or protective in heart disease. NFATc3 is a transcriptional mediator of cardiac stress signalling that promotes pathological hypertrophy[72]. NFATc3 was recently shown to induce miR-23a expression in cardiomyocytes, and antagomir-based knockdown of miR-23a in mice abrogates isoproterenol-induced cardiac hypertrophy[73]. Conversely, acute knockdown of miR-133 was shown to induce pathological cardiac hypertrophy in mice[70], suggesting a potential cardioprotective role for endogenous miR-133. However, these findings contrast with the phenotype of *Mir133*-null mice, which undergo a normal hypertrophic response[46], highlighting the difference between pharmacological modulation of miRNA expression and genetic deletion studies.

The miR-29 family, which is downregulated after myocardial infarction, inhibits the expression of several collagens and extracellular matrix proteins, thereby contributing to scar formation and fibrosis[14]. Similarly, the miR-199 family is rapidly downregulated in cardiac myocytes under hypoxic conditions, relieving the repression of sirtuin 1 and hypoxia-inducible factor 1-α in a model of hypoxia preconditioning[74].

The miRNA that repeatedly shows dynamic regulation after cellular stress is miR-21, which was shown to promote cardiac hypertrophy and fibrosis in response to pressure overload. Knockdown of miR-21 with a cholesterol-modified antagomir attenuated cardiac remodelling after thoracic aortic constriction[68]. This response was attributed to the derepression of the protein sprouty, which negatively regulates the pro-fibrotic extracellular signal-regulated kinase–mitogen-activated protein kinase (ERK–MAPK) cascade in cardiac fibroblasts[68]. Paradoxically, however, neither genetic deletion nor tiny locked-nucleic-acid (LNA)-mediated knockdown of miR-21 in mice alters fibrosis or hypertrophy in response to thoracic aortic constriction or other cardiac stresses[75]. The contrasting conclusions of these studies emphasize the gaps in our understanding of the mechanisms of miRNA action and oligonucleotide-based targeting strategies for their inhibition.

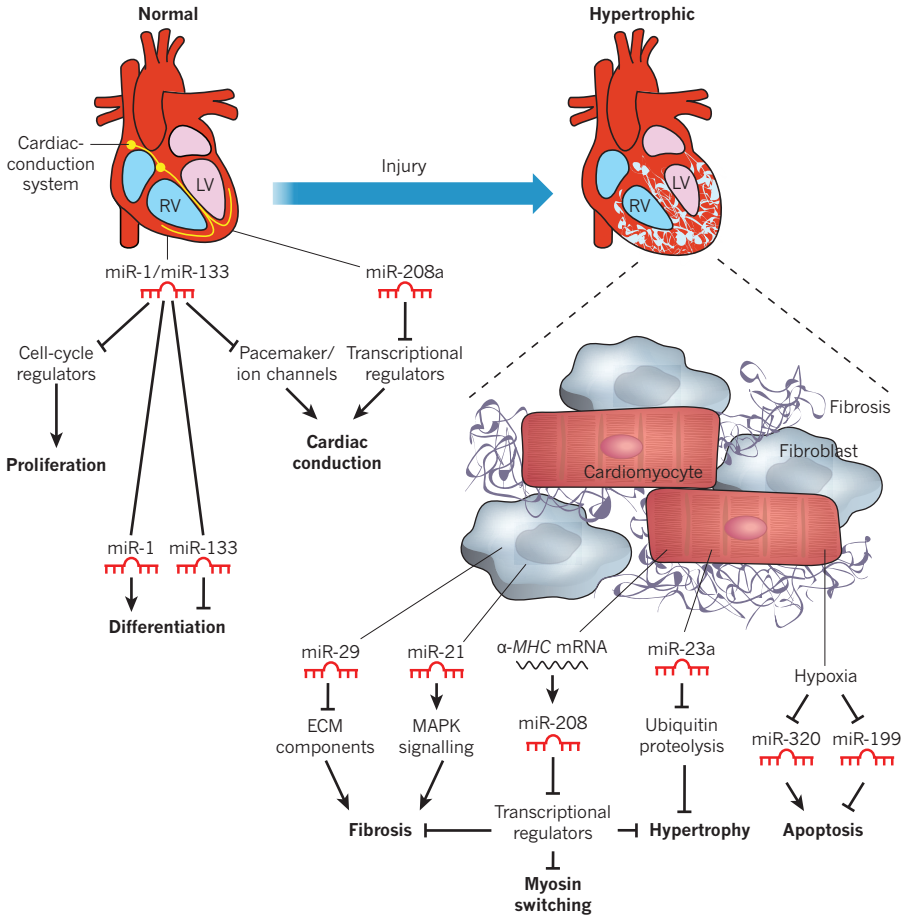


Figure 3 | Functional role of miRNAs in the normal and diseased heart. A normal and a hypertrophic heart are shown in schematic form, depicting miRNAs that contribute to normal function or pathological remodelling. The expression of selected miRNAs within the heart is shown, along with their corresponding functions. All arrows denote the normal action of each component or process. miR-1 and miR-133 are involved in the development of a normal heart (left) by regulating proliferation, differentiation and cardiac conduction. For example, proliferation is promoted by cell-cycle regulators, but miR-1 and miR-133 block these regulators, thus blocking proliferation. miR-208a also contributes to the regulation of the conduction system. After cardiac injury (right), various miRNAs contribute to pathological remodelling and the progression to heart failure. miR-29 and miR-21 block and promote cardiac fibrosis, respectively. miR-29 blocks fibrosis by inhibiting the expression of ECM components, whereas miR-21 promotes fibrosis by stimulating mitogen-activated protein kinase (MAPK) signalling. miR-208 controls myosin isoform switching, cardiac hypertrophy and fibrosis. miR-23a promotes cardiac hypertrophy by inhibiting ubiquitin proteolysis, which itself inhibits hypertrophy. Hypoxia results in the repression of miR-320 and miR-199, which promote and block apoptosis, respectively. ECM, extracellular matrix; LV, left ventricle; MHC, myosin heavy chain; RV, right ventricle.

Roles of miRNAs in vascular disease

The vessel wall is composed of endothelial cells and SMCs that must maintain a sealed barrier yet allow the exchange of oxygen and nutrients with adjacent tissues. Vessels can respond to injury or changes in the environment by undergoing phenotypic changes that promote endothelial cell migration or fragility, as well as SMC de-differentiation, proliferation and migration. Numerous miRNAs show marked alterations in expression during vascular injury and disease, and expression signatures have now been correlated with pathologies such as ischaemia, tumour angiogenesis, atherosclerosis and a proliferative thickening and obstruction of the vessel known as restenosis[15]. Some miRNAs have been shown to have causal roles in these disorders (Fig. 4).

Angiogenesis is a process of endothelial cell proliferation and vascular tube sprouting that is promoted in adulthood by various stimuli, including tumour growth, retinal damage and ischaemia. miR-21 can influence the function and migration of angiogenic progenitor cells during coronary artery disease[76]. Likewise, ischaemia-induced angiogenesis in adult tissues can be promoted or inhibited by antisense oligonucleotides directed against miR-92a (ref. 77) or miR-126 (ref. 78), respectively. miR-126 may also influence susceptibility to atherosclerosis, through the modification of endothelial cell function[79,80].

Vessel injury, instigated by diverse factors such as atherosclerosis, hypertension and damage due to a mechanical stenting, results in SMC phenotypic changes indicative of a de-differentiated state. Such SMCs become proliferative and migratory, entering the vessel lumen and causing restenosis. Recent studies have implicated miRNAs as mediators of SMC phenotypic modulation and vessel remodelling. The expression of miR-21 and the miR-143/145 cluster are up- and downregulated, respectively, after mechanical injury of large vessels[15], and restoration of miR-21 and miR-145 to normal levels prevents restenosis[15,56,81].

miRNA mutations as the basis of disease

The pervasive influence of miRNAs on cardiovascular function and disease raises questions as to whether polymorphisms in miRNAs or their target sequences in mRNA transcripts affect human disease. Although mutations within the seed regions of evolutionarily conserved miRNAs are not common, single nucleotide polymorphisms (SNPs) within miRNA-binding sites in the 3′ UTRs of target mRNAs have been observed at a higher frequency[82,83]. A notable example is in the Texel breed of sheep, which develops extreme skeletal muscle hypertrophy owing to an SNP in the 3′ UTR of the mRNA encoding myostatin, a negative regulator of muscle growth[84]. This mutation creates a binding site for miR-1, resulting in repression of myostatin expression and unrestricted muscle growth. SNPs within potential miRNA-binding sites have also been identified in mRNAs associated with hypertension and cardiovascular disease[85]. SNPs in miRNAs or their targets that cause significant phenotypes seem unlikely to be a widespread occurrence, however, because of the substantial degeneracy allowed in miRNA–mRNA interactions and the redundant regulation of an individual mRNA by several unrelated miRNAs. It remains to be determined whether a causative link can be made between miRNA SNPs and human disease.

Clinical perspectives

Identifying the signature patterns of miRNAs associated with different cardiovascular disorders has opened up opportunities for miRNA diagnostics. miRNA profiling can discriminate between specific forms of heart disease[10,12], such as dilated cardiomyopathy, ischaemic cardiomyopathy and heart failure, and disease-associated miRNA expression patterns in failing human hearts can be normalized by the stabilization of cardiac output[11,86]. Recently, several miRNAs have been detected in plasma and reported to be diagnostic for heart failure and myocardial infarction[87–89]. Whether circulating miRNAs are functionally relevant

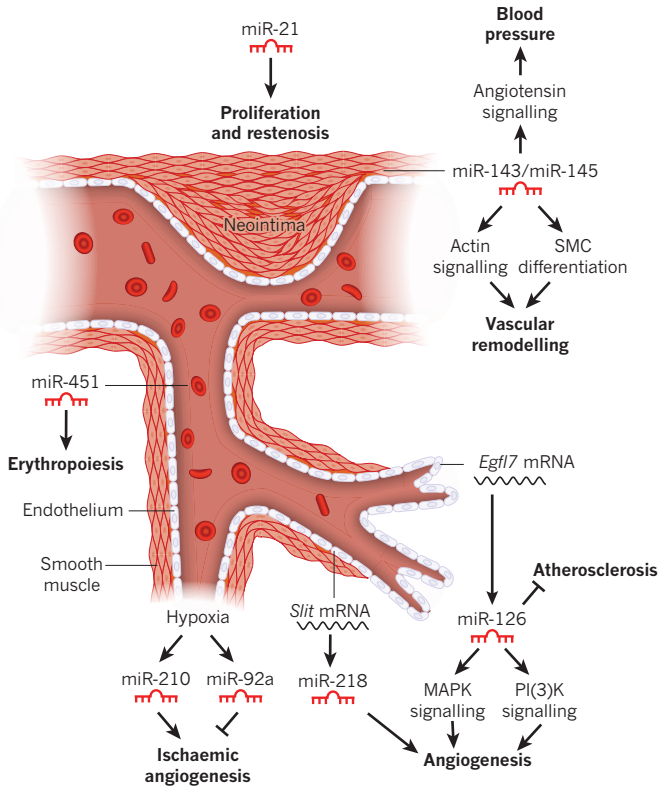


Figure 4 | Functional role of miRNAs in the vascular system. Blood vessel schematic showing the endothelial and smooth muscle layers, red blood cells and the proliferating SMCs of a neointimal lesion. The expression of select miRNAs is shown, along with their observed functional role. Hypoxia results in the activation of miR-210 and miR-92a, which promote and inhibit angiogenesis, respectively. miR-126, an endothelial-cell-enriched miRNA encoded by an intron of the *Egfl7* gene, modulates atherosclerosis and angiogenesis by regulating MAPK and PI(3)K signalling. Angiogenesis is also regulated by miR-218, which is encoded by an intron of the *Slit* genes. miR-143 and miR-145 are expressed in SMCs and control blood pressure and vascular tone, and contribute to vascular remodelling. miR-21 is induced in SMCs after vascular injury, and promotes proliferation and neointima formation. miR-451 regulates the proliferation and differentiation of erythroid cells.

or are simply released from injured tissues remains to be determined. However, the cellular secretion of particular miRNAs by exosomes suggests specificity in the process of miRNA secretion.

In contrast to many cellular mediators of disease, which are difficult (or impossible) to modulate therapeutically, it is unquestionable that drugs can target miRNAs. Thus, the involvement of miRNAs in almost every aspect of cardiovascular disease raises exciting possibilities for the therapeutic manipulation of miRNA-regulated processes. Therapies based on antimiRs or miRNA mimics are now being developed to repress pathological miRNAs or overexpress protective miRNAs, respectively. Indeed, antimiR-based studies demonstrating efficacy in non-human primates have already been reported[90,91] and have been advanced to human clinical trials.

The ability of individual miRNAs to modulate complex disease pathways through the targeting of several components of regulatory networks enables miRNAs to modulate tissue stress responses in a manner that is distinct from that of classical drugs. The multiplicity of miRNA targets also enables miRNAs to bypass mechanisms that render cells or tissues insensitive to certain drugs. For example, cells can develop insensitivity to single drugs through rare mutations in drug targets or desensitization of cell-surface receptors. Such mechanisms are unlikely to diminish sensitivity to miRNA inhibitors, which target several steps in a disease pathway.

Conversely, the targeting of large collections of mRNAs raises possibilities for off-target effects or even opposing effects of miRNAs in different tissues. Because the mechanistic basis of miRNA-based therapeutics is not clear, the possibility exists that modulating such a diverse set of target mRNAs will affect beneficial processes as well as the pathological condition. The heart takes up globally administered antimiR oligonucleotides less efficiently than the kidneys and liver, and the pharmacokinetics of miRNA-based therapies remain a hurdle. This issue may necessitate the development of new cardiovascular delivery systems for miRNA-based therapeutics, to limit uptake in healthy tissue. These methods would not be required for strategies involving knockdown of cardiac-specific miRNAs. Conjugation of antimiRs or miRNA mimics to homing molecules such as peptides, antibodies or other bioactive molecules might enrich uptake in cardiac tissue. This technology has not yet been successfully translated to the clinic, however, and other methods may improve tissue-specific uptake, such as direct administration by cardiac catheterization or a drug-coated stent.

Looking to the future

Despite recent advances in identifying miRNA contributions to cardiovascular development and disease, as well as in developing miRNA diagnostics and miRNA inhibitors, many gaps remain in our knowledge of miRNA-based regulation of gene expression in the normal and diseased heart and cardiovascular system. For example, the many potential target mRNAs for each miRNA pose significant challenges to the identification of those mRNAs that are relevant to a particular miRNA-regulated process. Compounding this difficulty is the apparent variability in miRNA function based on physiological context or cell type, making it necessary to define the potential disparate functions of individual miRNAs in different settings. Another important consideration is that combinatorial interactions between multiple miRNAs with common or coordinated target mRNAs are likely to have a major role in gene regulation and the control of physiological pathways. Thus, it will be crucial to identify sets of miRNAs acting cooperatively within the cardiovascular system. This information will be particularly relevant to the development of miRNA-based therapeutics, as cocktails of miRNA inhibitors may prove more efficacious than targeting a single miRNA.

With the current pace of advancements in deciphering the basic principles of miRNA action in cardiovascular development and disease, we foresee new therapeutic applications for the prevention and treatment of human pathologies based on miRNA biology in the relatively near future. ■

1. Hill, J. A. & Olson, E. N. Cardiac plasticity. *N. Engl. J. Med.* **358,** 1370–1380 (2008).
2. Hoffman, J. I. & Kaplan, S. The incidence of congenital heart disease. *J. Am. Coll. Cardiol.* **39,** 1890–1900 (2002).
3. Bruneau, B. G. The developmental genetics of congenital heart disease. *Nature* **451,** 943–948 (2008).
4. Cordes, K. R. & Srivastava, D. MicroRNA regulation of cardiovascular development. *Circ. Res.* **104,** 724–732 (2009).
5. Latronico, M. V. & Condorelli, G. MicroRNAs and cardiac pathology. *Nature Rev. Cardiol.* **6,** 419–429 (2009).
6. Small, E. M., Frost, R. J. & Olson, E. N. MicroRNAs add a new dimension to cardiovascular disease. *Circulation* **121,** 1022–1032 (2010).
7. van Rooij, E. & Olson, E. N. MicroRNAs: powerful new regulators of heart disease and provocative therapeutic targets. *J. Clin. Invest.* **117,** 2369–2376 (2007).
8. Liu, N. & Olson, E. N. MicroRNA regulatory networks in cardiovascular development. *Dev. Cell* **18,** 510–525 (2010).
9. Ikeda, S. *et al.* Altered microRNA expression in human heart disease. *Physiol. Genomics* **31,** 367–373 (2007).
10. van Rooij, E. *et al.* A signature pattern of stress–responsive microRNAs that can evoke cardiac hypertrophy and heart failure. *Proc. Natl Acad. Sci. USA* **103,** 18255–18260 (2006).
This important paper describes the dynamic regulation of miRNA expression during cardiac stress.
11. Matkovich, S. J. *et al.* Reciprocal regulation of myocardial microRNAs and messenger RNA in human cardiomyopathy and reversal of the microRNA signature by biomechanical support. *Circulation* **119,** 1263–1271 (2009).
12. Thum, T. *et al.* MicroRNAs in the human heart: a clue to fetal gene reprogramming in heart failure. *Circulation* **116,** 258–267 (2007).
13. Roy, S. *et al.* MicroRNA expression in response to murine myocardial infarction:

miR-21 regulates fibroblast metalloprotease-2 via phosphatase and tensin homologue. *Cardiovasc. Res.* **82,** 21–29 (2009).
14. van Rooij, E. *et al.* Dysregulation of microRNAs after myocardial infarction reveals a role of miR-29 in cardiac fibrosis. *Proc. Natl Acad. Sci. USA* **105,** 13027–13032 (2008).
15. Ji, R. *et al.* MicroRNA expression signature and antisense-mediated depletion reveal an essential role of microRNA in vascular neointimal lesion formation. *Circ. Res.* **100,** 1579–1588 (2007).
16. Xin, M. *et al.* MicroRNAs miR-143 and miR-145 modulate cytoskeletal dynamics and responsiveness of smooth muscle cells to injury. *Genes Dev.* **23,** 2166–2178 (2009).
17. Huang, Z. P., Neppl, R. L. & Wang, D. Z. MicroRNAs in cardiac remodeling and disease. *J. Cardiovasc. Transl. Res.* **3,** 212–218 (2010).
18. van Rooij, E., Marshall, W. S. & Olson, E. N. Toward microRNA-based therapeutics for heart disease: the sense in antisense. *Circ. Res.* **103,** 919–928 (2008).
19. Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136,** 215–233 (2009).
20. Filipowicz, W., Bhattacharyya, S. N. & Sonenberg, N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature Rev. Genet.* **9,** 102–114 (2008).
21. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116,** 281–297 (2004).
22. Lutter, D., Marr, C., Krumsiek, J., Lang, E. W. & Theis, F. J. Intronic microRNAs support their host genes by mediating synergistic and antagonistic regulatory effects. *BMC Genomics* **11**, 224 (2010).
23. Cao, G. *et al.* Intronic miR-301 feedback regulates its host gene, *ska2*, in A549 cells by targeting MEOX2 to affect ERK/CREB pathways. *Biochem. Biophys. Res. Commun.* **396,** 978–982 (2010).
24. Najafi-Shoushtari, S. H. *et al.* MicroRNA-33 and the SREBP host genes cooperate to control cholesterol homeostasis. *Science* **328,** 1566–1569 (2010).
25. Poliseno, L. *et al.* Identification of the *miR-106b~25* microRNA cluster as a proto-oncogenic *PTEN*-targeting intron that cooperates with its host gene *MCM7* in transformation. *Sci. Signal.* **3,** ra29 (2010).
26. Barik, S. An intronic microRNA silences genes that are functionally antagonistic to its host gene. *Nucleic Acids Res.* **36,** 5232–5241 (2008).
27. Alvarez-Saavedra, E. & Horvitz, H. R. Many families of *C. elegans* microRNAs are not essential for development or viability. *Curr. Biol.* **20,** 367–373 (2010).
28. Ambros, V. MicroRNAs: genetically sensitized worms reveal new secrets. *Curr. Biol.* **20,** R598–R600 (2010).
29. Brenner, J. L., Jasiewicz, K. L., Fahley, A. F., Kemp, B. J. & Abbott, A. L. Loss of individual microRNAs causes mutant phenotypes in sensitized genetic backgrounds in *C. elegans*. *Curr. Biol.* **20,** 1321–1325 (2010).
This paper suggests redundant and stress-responsive roles of miRNAs, through using miRNA mutants in *Dicer*-deficient *C. elegans*.
30. Zhao, Y. *et al.* Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking miRNA-1-2. *Cell* **129,** 303–317 (2007).
This paper demonstrates an important role for an miRNA in heart development by genetic deletion in mice.
31. Yang, B. *et al.* The muscle-specific microRNA miR-1 regulates cardiac arrhythmogenic potential by targeting *GJA1* and *KCNJ2*. *Nature Med.* **13,** 486–491 (2007).
32. Luo, X. *et al.* Down-regulation of miR-1/miR-133 contributes to re-expression of pacemaker channel genes *HCN2* and *HCN4* in hypertrophic heart. *J. Biol. Chem.* **283,** 20045–20052 (2008).
33. Small, E. M. *et al.* Regulation of PI3-kinase/Akt signalling by muscle-enriched microRNA-486. *Proc. Natl Acad. Sci. USA* **107,** 4218–4223 (2010).
34. Xu, N., Papagiannakopoulos, T., Pan, G., Thomson, J. A. & Kosik, K. S. MicroRNA-145 regulates OCT4, SOX2, and KLF4 and represses pluripotency in human embryonic stem cells. *Cell* **137,** 647–658 (2009).
35. Choi, W. Y., Giraldez, A. J. & Schier, A. F. Target protectors reveal dampening and balancing of Nodal agonist and antagonist by miR-430. *Science* **318,** 271–274 (2007).
36. Xiao, J. *et al.* Novel approaches for gene-specific interference via manipulating actions of microRNAs: examination on the pacemaker channel genes *HCN2* and *HCN4*. *J. Cell. Physiol.* **212,** 285–292 (2007).
37. Brown, B. D. & Naldini, L. Exploiting and antagonizing microRNA regulation for therapeutic and experimental applications. *Nature Rev. Genet.* **10,** 578–585 (2009).
38. Ebert, M. S., Neilson, J. R. & Sharp, P. A. MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nature Methods* **4,** 721–726 (2007).
39. Chen, J. F. *et al.* Targeted deletion of Dicer in the heart leads to dilated cardiomyopathy and heart failure. *Proc. Natl Acad. Sci. USA* **105,** 2111–2116 (2008).
40. Albinsson, S. *et al.* MicroRNAs are necessary for vascular smooth muscle growth, differentiation, and function. *Arterioscler. Thromb. Vasc. Biol.* **30,** 1118–1126 (2010).
41. Rao, P. K. *et al.* Loss of cardiac microRNA-mediated regulation leads to dilated cardiomyopathy and heart failure. *Circ. Res.* **105,** 585–594 (2009).
Deep sequencing showed that the 18 most abundant cardiac miRNAs account for more than 90% of all miRNAs in the heart.
42. Brown, B. D. *et al.* Endogenous microRNA can be broadly exploited to regulate transgene expression according to tissue, lineage and differentiation state. *Nature Biotechnol.* **25,** 1457–1467 (2007).
43. Liu, N. *et al.* An intragenic MEF2-dependent enhancer directs muscle-specific expression of microRNAs 1 and 133. *Proc. Natl Acad. Sci. USA* **104,** 20844–20849 (2007).
44. Zhao, Y., Samal, E. & Srivastava, D. Serum response factor regulates a muscle-specific microRNA that targets *Hand2* during cardiogenesis. *Nature* **436,** 214–220 (2005).
45. Ivey, K. N. *et al.* MicroRNA regulation of cell lineages in mouse and human embryonic stem cells. *Cell Stem Cell* **2,** 219–229 (2008).
46. Liu, N. *et al.* microRNA-133a regulates cardiomyocyte proliferation and suppresses smooth muscle gene expression in the heart. *Genes Dev.* **22,** 3242–3254 (2008).
47. Deacon, D. C. *et al.* The miR-143–adducin3 pathway is essential for cardiac chamber morphogenesis. *Development* **137,** 1887–1896 (2010).
48. Morton, S. U. *et al.* microRNA-138 modulates cardiac patterning during embryonic development. *Proc. Natl Acad. Sci. USA* **105,** 17830–17835 (2008).
49. Schmidt, M. *et al.* EGFL7 regulates the collective migration of endothelial cells by restricting their spatial distribution. *Development* **134,** 2913–2923 (2007).
50. Nicoli, S. *et al.* MicroRNA-mediated integration of haemodynamics and Vegf signalling during angiogenesis. *Nature* **464,** 1196–1200 (2010).
51. Kuhnert, F. *et al.* Attribution of vascular phenotypes of the murine *Egfl7* locus to the microRNA *miR-126*. *Development* **135,** 3989–3993 (2008).
52. Wang, S. *et al.* The endothelial-specific microRNA miR-126 governs vascular integrity and angiogenesis. *Dev. Cell* **15,** 261–271 (2008).
53. Fish, J. E. *et al.* miR-126 regulates angiogenic signalling and vascular integrity. *Dev. Cell* **15,** 272–284 (2008).
References 52 and 53 show a crucial role for miR-126 in angiogenesis.
54. Small, E. M., Sutherland, L. B., Rajagopalan, R., Wang, S. & Olson, E. N. MicroRNA-218 regulates vascular patterning by modulation of Slit–Robo signaling. *Circ. Res.* **107,** 1336–1344 (2010).
55. Cordes, K. R. *et al.* miR-145 and miR-143 regulate smooth muscle cell fate and plasticity. *Nature* **460,** 705–710 (2009).
56. Elia, L. *et al.* The knockout of miR-143 and -145 alters smooth muscle cell maintenance and vascular homeostasis in mice: correlates with human disease. *Cell Death Differ.* **16,** 1590–1598 (2009).
57. Boettger, T. *et al.* Acquisition of the contractile phenotype by murine arterial smooth muscle cells depends on the *Mir143/145* gene cluster. *J. Clin. Invest.* **119,** 2634–2647 (2009).
58. Zhao, G., Yu, D. & Weiss, M. J. MicroRNAs in erythropoiesis. *Curr. Opin. Hematol.* **17**, 155–162 (2010).
59. Georgantas, R. W. III *et al.* $CD34^+$ hematopoietic stem-progenitor cell microRNA expression and function: a circuit diagram of differentiation control. *Proc. Natl Acad. Sci. USA* **104,** 2750–2755 (2007).
60. O'Carroll, D. *et al.* A Slicer-independent role for Argonaute 2 in hematopoiesis and the microRNA pathway. *Genes Dev.* **21,** 1999–2004 (2007).
61. Lu, J. *et al.* MicroRNA-mediated control of cell fate in megakaryocyte–erythrocyte progenitors. *Dev. Cell* **14,** 843–853 (2008).
62. Wang, Q. *et al.* MicroRNA miR-24 inhibits erythropoiesis by targeting activin type I receptor ALK4. *Blood* **111,** 588–595 (2008).
63. Rasmussen, K. D. *et al.* The miR-144/451 locus is required for erythroid homeostasis. *J. Exp. Med.* **207,** 1351–1358 (2010).
64. Johnnidis, J. B. *et al.* Regulation of progenitor cell proliferation and granulocyte function by microRNA-223. *Nature* **451,** 1125–1129 (2008).
65. Patrick, D. M. *et al.* Defective erythroid differentiation in miR-451 mutant mice mediated by 14-3-3ζ. *Genes Dev.* **24,** 1614–1619 (2010).
66. Yu, D. *et al.* miR-451 protects against erythroid oxidant stress by repressing 14-3-3ζ. *Genes Dev.* **24,** 1620–1633 (2010).
References 65 and 66 show that miR-451 is required for proper erythroid differentiation, and suggest a potential therapeutic application for targeting miR-451 for degradation.
67. Callis, T. E. *et al.* MicroRNA-208a is a regulator of cardiac hypertrophy and conduction in mice. *J. Clin. Invest.* **119,** 2772–2786 (2009).
68. Thum, T. *et al.* MicroRNA-21 contributes to myocardial disease by stimulating MAP kinase signalling in fibroblasts. *Nature* **456,** 980–984 (2008).
This paper demonstrates an important role for miR-21 in cardiac remodelling using antagomir-mediated knockdown in mice.
69. van Rooij, E. *et al.* Control of stress-dependent cardiac growth and gene expression by a microRNA. *Science* **316,** 575–579 (2007).
The first paper to show a role for an miRNA, miR-208a, in the control of cardiac remodelling, using a genetic knockout.
70. Care, A. *et al.* MicroRNA-133 controls cardiac hypertrophy. *Nature Med.* **13,** 613–618 (2007).
71. van Rooij, E. *et al.* A family of microRNAs encoded by myosin genes governs myosin expression and muscle performance. *Dev. Cell* **17,** 662–673 (2009).
72. Molkentin, J. D. *et al.* A calcineurin-dependent transcriptional pathway for cardiac hypertrophy. *Cell* **93,** 215–228 (1998).
73. Lin, Z. *et al.* miR-23a functions downstream of NFATc3 to regulate cardiac hypertrophy. *Proc. Natl Acad. Sci. USA* **106,** 12103–12108 (2009).
74. Rane, S. *et al.* Downregulation of miR-199a derepresses hypoxia-inducible factor-1α and Sirtuin 1 and recapitulates hypoxia preconditioning in cardiac myocytes. *Circ. Res.* **104,** 879–886 (2009).
75. Patrick, D. M. *et al.* Stress-dependent cardiac remodeling occurs in the absence of microRNA-21 in mice. *J. Clin. Invest.* **120,** 3912–3916 (2010).
76. Fleissner, F. *et al.* Asymmetric dimethylarginine impairs angiogenic progenitor cell function in patients with coronary artery disease through a microRNA-21-dependent mechanism. *Circ. Res.* **107,** 138–143 (2010).
77. Bonauer, A. *et al.* MicroRNA-92a controls angiogenesis and functional recovery of ischemic tissues in mice. *Science* **324,** 1710–1713 (2009).
78. van Solingen, C. *et al.* Antagomir-mediated silencing of endothelial cell specific

microRNA-126 impairs ischemia-induced angiogenesis. *J. Cell. Mol. Med.* **13,** 1577–1585 (2009).
79. Zernecke, A. *et al.* Delivery of microRNA-126 by apoptotic bodies induces CXCL12-dependent vascular protection. *Sci. Signal.* **2,** ra81 (2009).
80. Harris, T. A., Yamakuchi, M., Ferlito, M., Mendell, J. T. & Lowenstein, C. J. MicroRNA-126 regulates endothelial expression of vascular cell adhesion molecule 1. *Proc. Natl Acad. Sci. USA* **105,** 1516–1521 (2008).
81. Cheng, Y. *et al.* MicroRNA-145, a novel smooth muscle cell phenotypic marker and modulator, controls vascular neointimal lesion formation. *Circ. Res.* **105,** 158–166 (2009).
82. Saunders, M. A., Liang, H. & Li, W. H. Human polymorphism at microRNAs and microRNA target sites. *Proc. Natl Acad. Sci. USA* **104,** 3300–3305 (2007).
83. Chen, K. & Rajewsky, N. Natural selection on human microRNA binding sites inferred from SNP data. *Nature Genet.* **38,** 1452–1456 (2006).
84. Clop, A. *et al.* A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nature Genet.* **38,** 813–818 (2006).
85. Sethupathy, P. *et al.* Human microRNA-155 on chromosome 21 differentially interacts with its polymorphic target in the *AGTR1* 3' untranslated region: a mechanism for functional single-nucleotide polymorphisms related to phenotypes. *Am. J. Hum. Genet.* **81,** 405–413 (2007).
86. Schipper, M. E., van Kuik, J., de Jonge, N., Dullens, H. F. & de Weger, R. A. Changes in regulatory microRNA expression in myocardium of heart failure patients on left ventricular assist device support. *J. Heart Lung Transplant.* **27,** 1282–1285 (2008).
87. Voellenkle, C. *et al.* MicroRNA signatures in peripheral blood mononuclear cells of chronic heart failure patients. *Physiol. Genomics* **42,** 420–426 (2010).
88. Fichtlscherer, S. *et al.* Circulating microRNAs in patients with coronary artery disease. *Circ. Res.* **107,** 677–684 (2010).
89. Ji, X. *et al.* Plasma miR-208 as a biomarker of myocardial injury. *Clin. Chem.* **55,** 1944–1949 (2009).
90. Elmen, J. *et al.* LNA-mediated microRNA silencing in non-human primates. *Nature* **452,** 896–899 (2008).
91. Lanford, R. E. *et al.* Therapeutic silencing of microRNA-122 in primates with chronic hepatitis C virus infection. *Science* **327,** 198–201 (2010).
The first report of therapeutically targeting an miRNA for the treatment of a disease in non-human primates.

Acknowledgements We apologize to all colleagues whose work could not be cited owing to space restrictions. We thank J. Cabrera for artwork and J. Brown for editorial assistance. E.N.O. was supported by grants from the National Institutes of Health, the Donald W. Reynolds Center for Clinical Cardiovascular Research, the Robert A. Welch Foundation, the Fondation Leducq's Transatlantic Network for Excellence in Cardiovascular Research Program, the American Heart Association and the Jon Holden DeHaan Foundation. E.M.S. was supported by a scientist development grant from the American Heart Association.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence should be addressed to E.N.O. (eric.olson@utsouthwestern.edu).

REVIEW

doi:10.1038/nature09784

The Polycomb complex PRC2 and its mark in life

Raphaël Margueron[1,2,3] & Danny Reinberg[4]

Polycomb group proteins maintain the gene-expression pattern of different cells that is set during early development by regulating chromatin structure. In mammals, two main Polycomb group complexes exist — Polycomb repressive complex 1 (PRC1) and 2 (PRC2). PRC1 compacts chromatin and catalyses the monoubiquitylation of histone H2A. PRC2 also contributes to chromatin compaction, and catalyses the methylation of histone H3 at lysine 27. PRC2 is involved in various biological processes, including differentiation, maintaining cell identity and proliferation, and stem-cell plasticity. Recent studies of PRC2 have expanded our perspectives on its function and regulation, and uncovered a role for non-coding RNA in the recruitment of PRC2 to target genes.

The term *Polycomb* (*Pc*) initially referred to a *Drosophila* mutant that displayed improper body segmentation[1]. It was suggested that *Polycomb* encodes a negative regulator of the homeotic genes that are required for segmentation[2]. The Polycomb group (PcG) now defines a set of genes characterized by mutations that result in similar phenotypes to those of *Polycomb*. The crucial role of PcG proteins during development is highlighted by early embryonic lethality in mice after the deletion of genes encoding some of these proteins (*Eed*, *Ezh2* (also known as *Enx-1*), *Suz12* and *Ring1B* (*Rnf2*)). The antagonistic activities of the PcG and the trithorax families of proteins culminate in the maintenance, throughout development and adulthood, of the appropriate patterns of homeotic gene expression in a spatially defined manner[3]. PcG proteins are found in several families of multiprotein complexes, including the Polycomb repressive complexes PRC1 and PRC2 (Fig. 1). Two other PcG complexes were characterized in *Drosophila*, PHO-repressive complex (PhoRC) and Polycomb repressive deubiquitinase (PR-DUB), and their components have orthologues in mammals; however, the conservation of their functions has not yet been addressed[4–6].

Polycomb-mediated gene silencing is thought to rely mostly on the regulation of chromatin structure, in part through post-translational modification (PTM) of histones. Hence, the PRC2 complex is responsible for the methylation (di- and tri-) of Lys 27 of histone H3 (H3K27me2/3)[3,6] through its enzymatic subunits EZH1 and EZH2, whereas the PRC1 complex monoubiquitylates Lys 119 of histone H2A (H2AK119ub) via the ubiquitin ligases RING1A and RING1B (Fig. 1). In addition, some PRC1 complexes can regulate gene expression by compacting chromatin in a manner independent of enzymatic activity[7]. The PRC1 component Pc (known as CBX in mammals) binds specifically to the product of PRC2 catalysis, H3K27me3, leading to the hypothesis that PRC1 functions downstream of PRC2. Although this premise is still cited in the literature, its operational status is equivocal as there are genes targeted by PRC2 that lack H2AK119ub[8] and genes targeted by PRC1 in the absence of PRC2 (refs 9, 10). Nonetheless, PRC2 and PRC1 are often both required to maintain gene repression.

Owing to the pivotal role of PRC2 in the coordination of PcG protein function, the still partial characterization of PRC1 and PRC1-like complexes in mammals, and the existence of up-to-date reviews on PRC1 (refs 3, 6), this Review focuses primarily on mammalian PRC2. After considering PRC2 in terms of evolution, we evaluate the newly appreciated, variable composition of PRC2 and describe the function of its catalytic product and its localization. Finally, we discuss the biological roles of PRC2, and propose a model for its recruitment to target genes that involves non-coding RNA (ncRNA).

Evolution of PRC2

The core PRC2 complex, which is conserved from *Drosophila* to mammals, comprises four components: EZH1/2, SUZ12, EED and RbAp46/48 (also known as RBBP7/4). The composition of PRC1 complexes is more variable, with only two core common components — RING1A/B together with BMI1, MEL18 (PCGF2) or NSPC1 (PCGF1)[6,11] (Fig. 1). The presence of PRC2 in various unicellular eukaryotes led to the suggestion that it existed in the last common unicellular ancestor, but it was lost at times during evolution as exemplified by the cases of *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae*, in which PRC2 is absent[12]. Notably, the PRC2 components, in contrast to those of PRC1, underwent little duplication in mammals, with vertebrates containing two copies of enhancer of zeste homologue, EZH1 and EZH2 (ref. 12). *Drosophila* has two copies of the EED homologue, ESC and ESCL. Although ESC and ESCL are interchangeable[13], the same might not be true for EZH1 and EZH2, which have different expression patterns. EZH1 is present in both dividing and differentiated cells, whereas EZH2 is found only in actively dividing cells. Also, PRC2 complexes containing EZH1 (PRC2–EZH1) in lieu of EZH2 have low methyltransferase activity compared with PRC2–EZH2 (ref. 14). This indicates that PRC2–EZH2 establishes cellular H3K27me2/3 levels through its EZH2-mediated methyltransferase activity, and that PRC2–EZH1 restores H3K27me2/3 that could have been lost after histone exchange or through demethylase activity. Moreover, PRC2–EZH1 and –EZH2 have distinct chromatin-binding properties, as illustrated by the specific chromatin-compaction property of PRC2–EZH1 (ref. 14).

In contrast to mammals, PRC2 evolved towards a greater complexity in plants, with species such as *Arabidopsis thaliana* having up to 12 homologues of PRC2 components[15]. A homologue of the mammalian and *S. pombe* heterochromatin protein 1 (HP1) that binds to H3K9me3 also exists in plants and is denoted LHP1. LHP1 binds to H3K27me3 and interacts with the RING1 homologues AtRING1A and AtRING1B, suggesting the existence of a PRC1-like complex

[1]Institut Curie, 26 Rue d'Ulm, 75005 Paris, France. [2]CNRS UMR3215, 26 Rue d'Ulm, 75005 Paris, France. [3]INSERM U934, 26 Rue d'Ulm, 75005 Paris, France. [4]Howard Hughes Medical Institute, Department of Biochemistry, New York University School of Medicine, 522 First Ave, New York, New York 10016, USA.

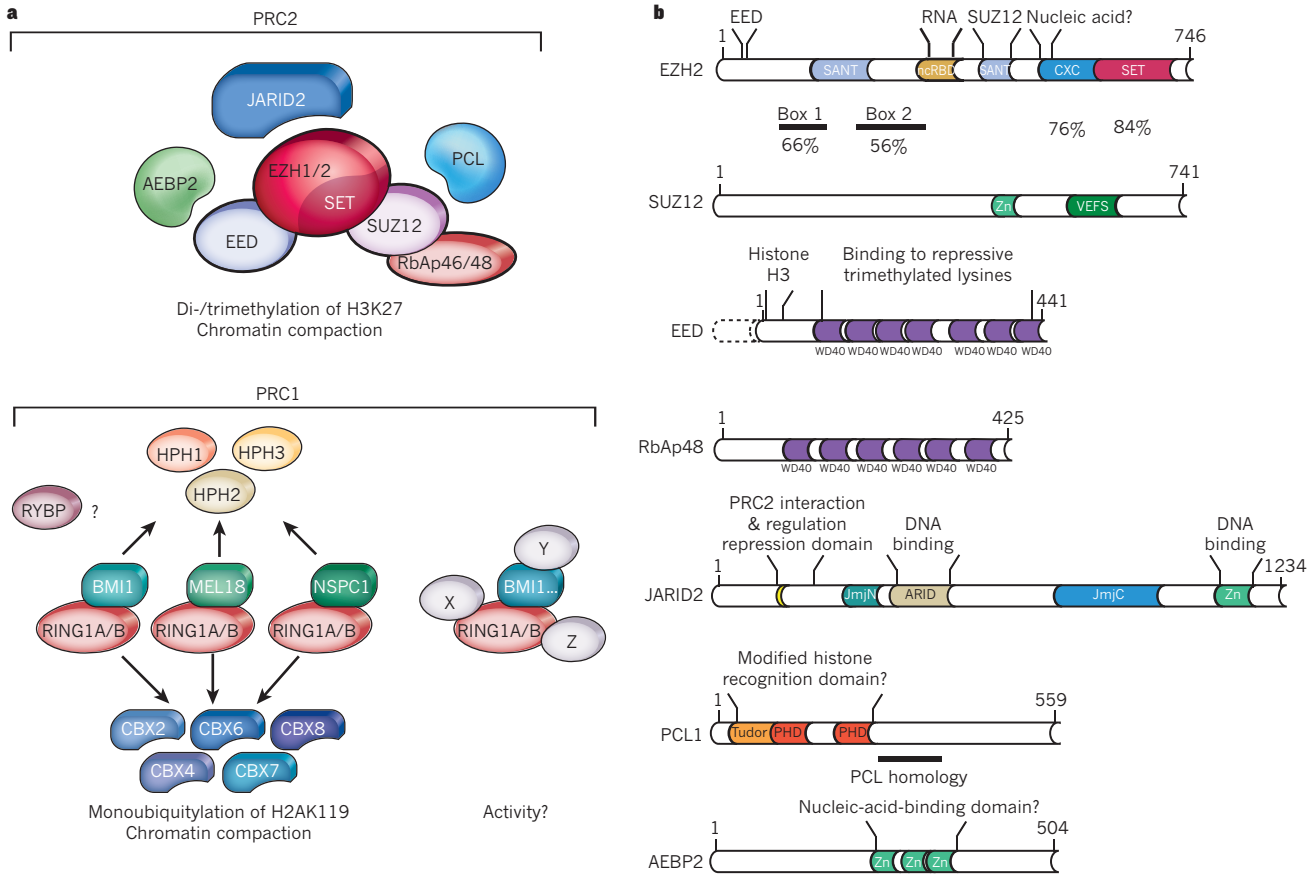


Figure 1 | The Polycomb complexes PRC1 and PRC2. **a**, Diagrams representing the composition of PRC2 and PRC1 are shown. In PRC1, the diagrams shown on the left correspond to the classical PRC1 complexes, whereas those on the right correspond to the so-called PRC1-like complexes. Owing to their homology with the *Drosophila* PSC protein, we assumed that the BMI1-, MEL18- and NSPC1-containing PRC1 complexes could compact chromatin. The 'pocket' shape of the CBX proteins represents the chromodomain that specifically recognized H3K9/27me3. HPH1, 2 and 3 denote human polyhomeotic homologue 1, 2 and 3. X, Y and Z denote various proteins such as SCMH1/2, FBXL10, E2F6 and JARID1D that could contribute to the formation of PRC1-like complexes, whose exact composition is still enigmatic. **b**, Characterized domains with potential functions are indicated for each PRC2 component. In EZH2, box 1 and 2 refer to domains based on sequence homology, and the numbers below the scheme indicate the percentage similarity between mouse and *Drosophila* homologues for the corresponding domain. CXC, cysteine-rich domain; ncRBD, non-coding-RNA-binding domain; SANT, SWI3, ADA2, N-CoR and TFIIIB DNA-binding domain; SET, Su(var)3-9, enhancer of zeste, trithorax domain; VEFS, conserved among VRN2–EMF2–FIS2–SU(Z)12; WD40, short ~40 amino acid motifs.

in plants[15]. Whereas EZH1 and EZH2 target the same genes and are expected to contribute to the same silencing pathway[16], the plant PRC2 complexes were reported to have distinct functions[15].

On the basis of these criteria, we speculate that PRC2 evolved from a function that was partially redundant with gene silencing through the H3K9me3 pathway, gaining a more specific role as multicellular organisms acquired specific cell lineages.

PRC2 comprises more than four components

The first PRC2 purifications led to the identification of the four components that are required for its enzymatic activity *in vitro*. It was recently shown that PRC2 contains several other polypeptides (Fig. 1) — AEBP2, PCLs and JARID2 — the functions of which are described below. Of note, other proteins transiently interact with PRC2 (for example, DNMTs, HDAC1 and SIRT1), but their effect on PRC2 function is unclear, and as such, they are not discussed further here.

AEBP2 is a zinc-finger protein that was identified as part of the PRC2 complex. It interacts with several PRC2 components to enhance its enzymatic activity[17], and co-localizes with PRC2 at some target genes[18]. AEBP2 was postulated to bind DNA with an apparently relaxed specificity[18].

PCL1, PCL2 and PCL3 (also known as PHF1, MTF2 and PHF19, respectively) are the three mammalian orthologues of *Drosophila* Polycomblike (PCL). They share the same protein motifs: a tudor domain, two plant homeodomain (PHD) finger proteins, a PCL extended domain and a carboxy-terminal domain tail[19] (Fig. 1). PCL proteins interact with PRC2 through EZH2, and to some extent through SUZ12 and the histone chaperones RbAp46 and RbAp48 (ref. 20). Genome-wide studies showed that PCL2 co-occupied PRC2 target genes[21,22]. Various functions have been attributed to PCLs, from the regulation of PRC2 enzymatic activity[20,23] to the gene recruitment of PRC2 (refs 21, 24). Mammalian PCLs are expressed in a tissue-specific manner[21], and this redundancy could explain apparent discrepancies between studies. The phenotypes associated with PCL mutation in *Drosophila* and *Xenopus*, and the co-localization and interaction of PCLs and PRC2, point to PCL proteins having a crucial role in PRC2 function. Understanding the underlying molecular mechanisms will probably require a detailed understanding of how PCLs interact with chromatin.

JARID2 is the founding member of the Jumonji family of proteins that catalyses the demethylation of histone proteins, yet it lacks the key residues necessary for cofactor binding and is devoid of enzymatic activity. Its deletion in mice results in severe defects in cardiovascular and liver development[25]. The C-terminal half of JARID2 contains some conserved regions such as the ARID domain (a potential DNA-binding domain), the JmjC and JmjN domains, and a zinc finger (Fig. 1). JARID2 was identified as a PRC2 component, and biochemical studies have

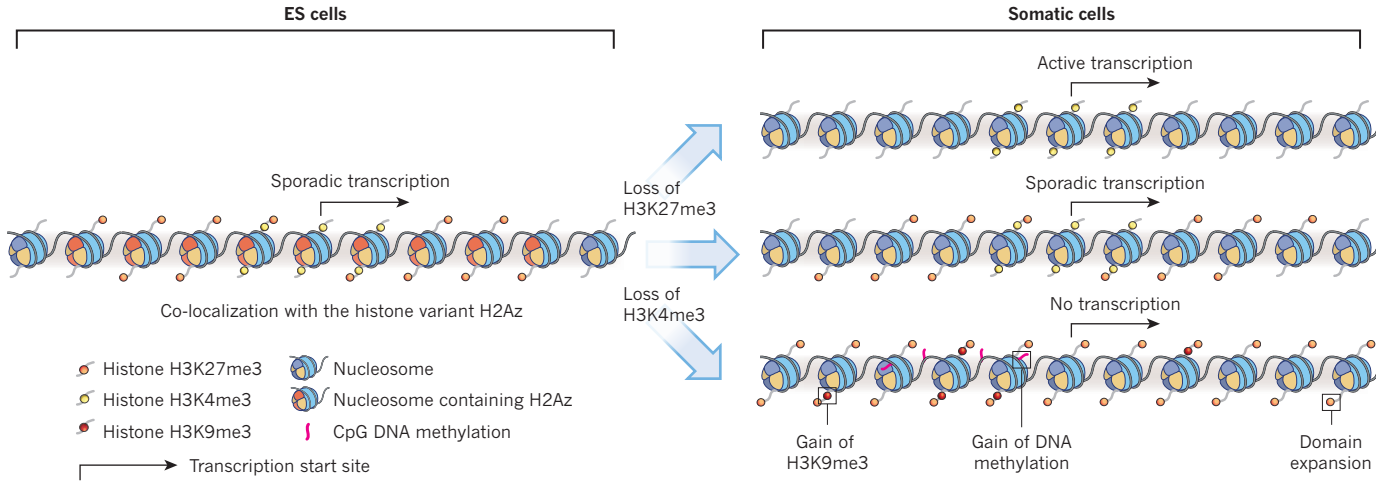


Figure 2 | Chromatin properties at PRC2 target genes in ES cells and differentiated somatic cells. Schematic representation of chromatin at PcG target genes as a function of ES-cell differentiation. In ES cells, most PcG targets are methylated on both H3K4 and H3K27 and co-localize with the histone variant H2Az. During differentiation, H2Az is removed, and some bivalent domains are resolved. For example, genes that are actively transcribed lose H3K27me3. A substantial proportion of PcG targets that retain H3K27me3 but lose H3K4me3 are targeted by other silencing pathways such as DNA methylation or H3K9 trimethylation.

demonstrated its interaction with EZH2 (refs 22, 26–29). Genome-wide studies showed a large overlap between PRC2 and JARID2 target genes[22,26–29], and the recruitment of JARID2 and PRC2 seems to be partially interdependent. Surprisingly, although PRC2 recruitment is impaired after JARID2 inactivation, H3K27me3 levels are only modestly affected[27,29]. This observation led to the conclusion that JARID2 is an inhibitor of PRC2 enzymatic activity[26,27]. However, further characterization of the catalytic activity of PRC2 indicated that JARID2 enhances PRC2 activity under defined biochemical conditions (ref. 22 and J. Son, R.M. and D.R., unpublished data). Furthermore, JARID2 is able to bind to DNA with a slight bias towards (G+C)-rich sequences[22]. This observation correlates with the reported sequence composition of PRC2 target genes[8], and is consistent with a potential function for JARID2 in PRC2 recruitment.

Studies of these three PRC2 components have given rise to apparent discrepancies. Considering that the factors are not strictly required for PRC2 enzymatic activity *in vitro*, it is perhaps unsurprising that their inactivation would result in milder consequences than the inactivation of a core PRC2 component. Nonetheless, these factors are necessary for optimum PRC2 activity, and the regulation of PRC2 recruitment and its enzymatic activity are tightly connected. We propose that PRC2 functions as a holoenzyme (Fig. 1), with the additive contribution of each of its components being required for maximum activity.

H3K27 methylation

It remains to be determined whether the composition of PRC2 is modified as a consequence of, or during the processes of, tumorigenesis, development and/or maintenance of adult tissue specificity. Thus far, despite the different proteins found associated with the core complex, its integrity remains intact such that all PRC2 complexes containing either EZH1 or EZH2 catalyse H3K27 methylation. Lysine can be mono-, di- or trimethylated, with each methylation level likely to be functionally distinct. Methylation of H3K27 is processive (H3K27me3 results from monomethylation of H3K27me2), and H3K27me3 is a stable mark[30]. Methylated H3K27 is very abundant, with roughly 50% of the H3 histone being dimethylated, 15% trimethylated and 15% monomethylated in embryonic stem (ES) cells[31]. Although the Pc component of PRC1 binds to H3K27me2 and -me3 through its chromodomain *in vitro*, it seems specific for H3K27me3 *in vivo*, and H3K27me2 seems to be of limited importance for maintenance of gene repression[23]. We previously proposed that H3K27me2 is an important intermediary PRC2 product, because it not only constitutes the substrate for subsequent H3K27me3 formation, but might also prevent H3K27 from being acetylated. Acetylated H3K27 is thought to be antagonistic to PcG-mediated silencing and is enriched in the absence of PRC2 (ref. 32).

With the exception of a viral protein, PRC2 has the only enzymatic activity found thus far that both di- and trimethylates H3K27. These methyl marks are associated with facultative heterochromatin — a subdivision of heterochromatin that is regulated in a developmental-specific manner[33]. The monomethylated version of H3K27 is associated with constitutive heterochromatin — a more stably silent, gene-poor region of chromatin — but its enrichment through the gene body is correlated with actively transcribed genes[34]. Exactly how H3K27me1 arises is still controversial. In plants, H3K27 is monomethylated by two enzymes, ATXR5 and ATXR6, which are distinct from PRC2 and not conserved in mammals[35]. But H3K27me1 in mammals is still detected in cells bearing non-functional PRC2 (refs 10, 36). We speculate that in mammals, H3K27me1 is placed by an enzymatic activity distinct from that of PRC2, and that the presence of H3K27me1 in actively transcribed genes could arise from the demethylation of H3K27me2/3 by the demethylases UTX or JMJD3 (ref. 37). Whether these demethylases can function on H3K27me1 *in vivo* is an open question.

In general, histone PTMs regulate biological processes either by altering chromatin structure (by loosening the DNA–histone interaction) or by contributing to the recruitment of further regulatory factors. Thus far, H3K27me3 has been implicated in only the latter mechanism of action, suggesting that other factors such as PRC1 are required to maintain gene repression. H3K27me3 might also indirectly regulate transcription by sterically preventing proteins from binding to chromatin. Enrichment of H3K27me3 correlates with gene silencing[38], and this observation is supported by the finding that H3K27me3 and H3K36me3, a mark that is linked to transcription elongation, have distinct localizations[39]. RNA polymerase (Pol) II that is phosphorylated at Ser 5 of its C-terminal domain is present in a substantial fraction of H3K27me3-enriched promoters[40], and low transcript levels are detected[41], leading to the suggestion that RNA Pol II could be paused at PcG-targeted genes[40]. Indeed, several PcG-regulated genes in *Drosophila* and mammals can recruit the RNA Pol II transcription complex to their respective promoters and engage in early transcription, yet these polymerases encounter an early block to elongation. A recent study indicates that short transcripts that are generated after transcription and remain bound to a paused RNA Pol II could recruit PRC2 (ref. 42). If confirmed, this suggests that PRC2 and H3K27me3 can affect gene expression by controlling an

engaged RNA Pol II during promoter escape or elongation, rather than by regulating the initiation phase of transcription. A likely possibility is that PRC2 can repress transcription by different mechanisms, and this may be gene specific.

Genome-wide localization of PRC2 and H3K27me3

Several publications have reported the genome-wide localization of H3K27me3 in various cell lines and organisms, with some divergent results depending on the methodology used and the model analysed. A conservative estimation is that PRC2 targets represent at least 10% of the genes in ES cells[43]. PRC2 specifically resides at — and targets for H3K27me3 deposition — the *Hox* genes and numerous genes encoding other developmental regulators[44–46]. Interestingly, in human cancer cells, the PRC2 component SUZ12 is mainly enriched at the promoters of genes encoding glycoprotein and immunoglobulin-like proteins[47]. Further studies are required to determine whether this is a consequence of the genetic and epigenetic alterations of cancer cells or whether it is a reflection of the cancer-cell origin.

In *Drosophila*, domains enriched in H3K27me3 were found to cover large regions of the genome, usually exceeding 10 kilobases (kb)[48,49]. In mammals, two different types of binding pattern have been reported for PRC2 or H3K27me3: some very large domains of more than 100 kb such as those containing the *Hox* loci, and some smaller domains covering a few kilobases[41,45,47,50]. H3K27me3 enrichment seems to be centred around the transcription start site of promoters, but with a lower intensity at the start site itself[41,51] (Fig. 2). Some H3K27me3 is found at intergenic regions[34,41], and H3K27me3 is enriched in subtelomeric regions[52] and in long-terminal repeat retrotransposons[53].

To understand how PRC2 can maintain specific gene-expression patterns, the overall chromatin structure, in addition to H3K27me3 patterns, should be considered[54]. This issue has generated a great deal of attention in the context of ES-cell differentiation (Fig. 2). ES cells are characterized by a more open and flexible chromatin organization and a higher overall rate of transcription, which is thought to be important for pluripotency[55]. Notably, the H3K4me3 mark, often associated with active transcription, was present at most, if not all, PRC2-targeted genes in ES cells, forming the 'bivalent domain'[39,41,43,51,56]. Although this pattern was initially believed to be ES-cell specific[56], bivalent domains have been found in differentiated somatic cells, albeit at a lower frequency[39,43]; they were also found in zebrafish[57] but are rarely detected in *Drosophila*[58]. Another histone species with seemingly disparate functionality that co-localizes with PRC2 is the histone variant H2Az, which is usually associated with active genes (Fig. 2). Indeed, PRC2 and H2Az co-localize in undifferentiated ES cells, and their recruitment is interdependent[59]. The apparent contradiction in the presence of either H3K4me3 or H2Az with H3K27me3 at the promoters of silent genes in ES cells might reflect the necessary plasticity of these cells, but could also result in partial leakiness of gene silencing. That PRC2 and H2Az co-localize is consistent with the low levels of DNA methylation at PcG target genes in ES cells[60,61], given the evolutionarily conserved exclusivity shown by H2Az and DNA methylation[62].

After ES-cell differentiation, a substantial fraction of bivalent domains that lose H3K4me3 and H2Az do gain DNA methylation[43,60,61]. Notably, genes enriched in both H3K27me3 and H3K9me3, another mark associated with gene repression, are more abundant in human fetal lung fibroblasts (IMR90) than in human ES cells[50]. In this same study, the authors showed that H3K27me3 domains are more extended in IMR90 cells or $CD4^+$ T cells than in ES cells, and that H3K27me3 domain expansion correlates with more efficient transcriptional silencing[50]. Altogether, these results indicate that somatic cells reinforce gene silencing by increasing the length of H3K27me3 domains and, for a fraction of PRC2-targeted genes, by complementary silencing pathways (H3K27me3 together with H3K9me3 or DNA methylation). Not surprisingly, some pluripotency factors, the expression of which could be deleterious in differentiated cells, are silenced in this redundant fashion[50].

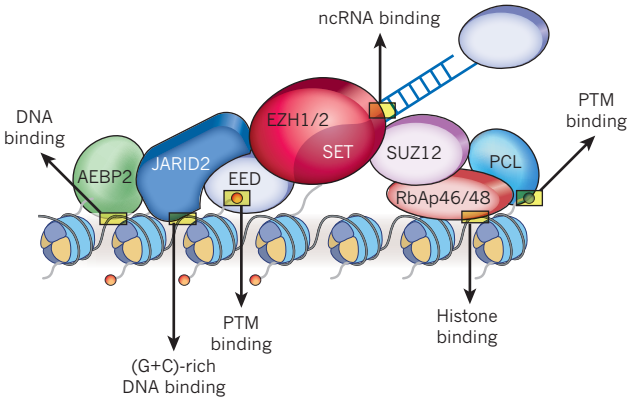


Figure 3 | The many interactions of PRC2 with chromatin. Schematic representation of the PRC2 holoenzyme at chromatin. Putative interactions with either DNA or histones that could explain PRC2 recruitment are highlighted.

PRC2 recruitment

Exactly how mammalian PRC2 is recruited to chromatin is not clear. In *Drosophila*, DNA sequences known as Polycomb response elements (PRE) are targets for PcG protein recruitment when inserted at exogenous loci[3,6]. Genetic experiments led to the identification of DNA-binding proteins that are required for PcG binding; however, genome-wide analysis showed that any one of these *trans*-acting factors only partially overlaps with PcG target genes. Instead, it is thought that a combination of these factors might be responsible for the recruitment of PcG proteins.

In mammals, PRC2-targeted sequences are highly enriched in C+G, most of them being classified as CpG islands, but these sequences alone do not indicate a consensus response element[8]. Recently, two publications identified a mammalian PRE on the basis of PcG complex recruitment in *Drosophila*[9,63]. Both reports suggested an important role for YY1, the mammalian orthologue of the *Drosophila* PRE DNA-binding protein PHO, as previously proposed[64]. RYBP, a protein that interacts with both YY1 and PRC1, was shown to be required for PRC1 and PRC2 recruitment[63]. Yet genome-wide analysis in mammals did not show a clear overlap between YY1 and PcG target genes[65]. Moreover, PRC2 is under-represented at YY1 response elements[8]. Hence, so far, there is no strong evidence for the involvement of transcription factors in the recruitment of PRC2 in mammals.

On the other hand, long ncRNAs are becoming recognized as important participants in PRC2 function. In mammals, X-chromosome inactivation initiates the expression of a 17-kb ncRNA, XIST, which coats the X chromosome *in cis*. Coating with XIST RNA leads to a marked alteration of chromatin structure characterized by a progressive heterochromatinization. The inactive X chromosome becomes methylated at H3K27 in an XIST-dependent manner[66]. The two long stem–loop structures formed by the A repeats present 5′ in the XIST RNA interact with PRC2 *in vitro*[67,68], although further regions of XIST are clearly involved because an XIST transcript in which the A repeats are deleted can still recruit PRC2 to the XIST RNA-coated X chromosome[69]. Similarly, the long ncRNA KCNQ1OT1 can mediate PRC2 spreading *in cis*, thereby maintaining the imprinted expression of the KCNQ1 domain[70]. Long ncRNA could also promote PRC2 binding *in trans* as shown for the RNA HOTAIR[71,72], the expression of which from the HOXC locus is associated with repression of 40 kb of the HOXD locus. Such mechanisms could be common to a large fraction of long ncRNAs[73]. In light of these results, ncRNA seems to be a strong candidate for PRC2 recruitment.

Considering this information, we propose a model in which the sum of relatively weak interactions or low energy steps that are established by each of the PRC2 holoenzyme components would function together to attain the necessary energy to recruit PRC2 (Fig. 3). This model predicts up to four steps, not necessarily consecutive, that result

in the successful recruitment of PRC2: (1) the interaction of JARID2 and AEBP2 with DNA[18,22]; (2) the interaction of the histone chaperones RbAp46 or 48 with histones H3 or H4 (ref. 74); (3) the interaction of EED with the product of PRC2 catalysis, H3K27me3 (ref. 75), and of PCLs with an unknown histone mark; and (4) the interaction of PRC2 components with long ncRNA. The resultant binding specificity could then be modulated by the variation in the composition of the PRC2 holoenzyme and PTMs of its components. Indeed, EZH2 was reported to be phosphorylated at Thr 350 (refs 76, 77), a modification that modulates PRC2 recruitment[76]. Consistent with the hypothesis that ncRNA will be a major player in the cell-specific recruitment of PRC2, phosphorylation of EZH2(T350) enhances its binding to ncRNA[77]. The large pool of long ncRNA may function, in part, to direct the complex to defined target genes. This targeting may not necessarily entail linear base pairing with target sequences, but instead the tertiary structure of the RNA may be key to specific target gene recognition. In this regard, the global contribution of HOTAIR to PRC2 targeting indicates that ncRNA may also regulate overall PRC2-binding properties to chromatin, either directly or by bridging it to other factors[72]. Hence, ncRNA could regulate the affinity of PRC2 to chromatin in a similar manner to the recently described case of the PRC1 component CBX7 (ref. 78). The chromodomain of CBX7 was reported to bind both H3K27me3 and the ncRNA ANRIL (also known as CDKN2B-AS1), and binding to one ligand can modulate the affinity for the other *in vitro*.

It is not yet clear whether the initial recruitment of PRC2 to a defined gene and the maintenance of its recruitment involve the same mechanisms. The PRC2 component EED can bind H3K27me3, and PRC2 enzymatic activity is stimulated by the presence of H3K27me3, thus generating a positive-feedback loop[75]. The importance of this mechanism is illustrated by the phenotype of *Drosophila* expressing EED point mutants that prevent its binding to H3K27me3 without altering PRC2 complex formation; this phenotype includes a global reduction in H3K27me2/3. Furthermore, given that some PcG proteins seem to stay bound to chromatin during replication[79] and that the same applies to PRC2 components during mitosis[80], PRC2 occupancy of chromatin may not necessitate its active recruitment to defined chromatin loci in all cases.

PRC2 pluripotency and differentiation

Two straightforward models could explain the maintenance of stem-cell pluripotency in the context of PRC2-mediated gene repression. Pluripotency is either lost after the expression of developmental regulators that promote differentiation, or lost when the expression of factors requisite for pluripotency are silenced (Fig. 4a). The first hypothesis is in keeping with the role of PRC2 in maintaining the repression of numerous developmental regulators in ES cells. This led to the suggestion that PRC2 is required for the maintenance of pluripotency[45]. However, later studies reported that ES cells in which a PRC2 component is inactivated could be kept undifferentiated. This finding draws our attention to the second model that posits the requisite repression of pluripotency-specific factors[16,36,81]. Indeed, in mouse ES cells, the inactivation of SUZ12, JARID2 or PCL2 was reported to be associated with an inefficient silencing of the pluripotency factors NANOG and OCT4 (also known as POU5F1)[21,22,29] (Fig. 4a). Furthermore, inactivation of the EZH2 homologue MES-2 extends the plasticity phase during embryonic development in *Caenorhabditis elegans*[82]. This observation probably results from the failure of the MES-2 mutant to repress genes that should only be expressed during a defined window of time in early development. Altogether, it seems that the sustained expression of pluripotency factors overtakes the aberrant expression of developmental regulators in PRC2-deficient ES cells.

By contrast, when ES cells are induced to differentiate, misregulation of developmental programs becomes more apparent. Hence, although *Eed*$^{-/-}$ ES cells (formally EED(L196P) but referred to as *Eed*$^{-/-}$ for simplicity) were unimpaired in their ability to contribute to all tissue lineages in chimaeric embryos[81], *Suz12*$^{-/-}$ ES cells fail to form a proper endodermal layer[36], and *Ezh2*$^{-/-}$ or *Eed*$^{-/-}$ ES cells display a severe defect in mesoendodermal lineage commitment[16]. This phenotype is not restricted to the deletion of PRC2 core components, because impaired differentiation was also reported in *Jarid2*$^{-/-}$ and *Pcl2*-knockdown ES cells[27–29]. Of note, in contrast to the mild phenotype that results from PRC2 deletion in ES cells, PRC1 inactivation (*Ring1a Ring1b* double knockout) leads to a proliferation defect, and ES cells cannot be maintained[83]. Furthermore, deletion of the PRC1 component RING1B in the context of *Eed*$^{-/-}$ ES cells worsens the differentiation defects[53]. These results indicate that PRC1 is not just a downstream effector of PRC2, but instead has distinct functions, and its recruitment is at least partially PRC2 independent.

On the basis of the example of ES-cell differentiation, we would expect PRC2 inactivation to prevent lineage commitment and terminal differentiation. Although PRC2 defects do prevent adipogenesis and lymphopoiesis[84,85], PRC2 inactivation also promotes differentiation during myogenesis and epidermis formation[86,87] (Fig. 4b). During B-cell maturation, EZH2 is required for V_HJ558 gene rearrangement, and the transition from pro-B to pre-B cells is altered in its absence[84]. In the case of epidermis, PRC2 inactivation leads to upregulation of epidermal genes mediated by the transcription factor AP1. Those genes are normally expressed at the late stage of differentiation[87]. Considering that only a small subset of its target genes is reactivated after PRC2 inactivation, it is likely that individual functions encoded by these genes dictate the global consequences for cell differentiation.

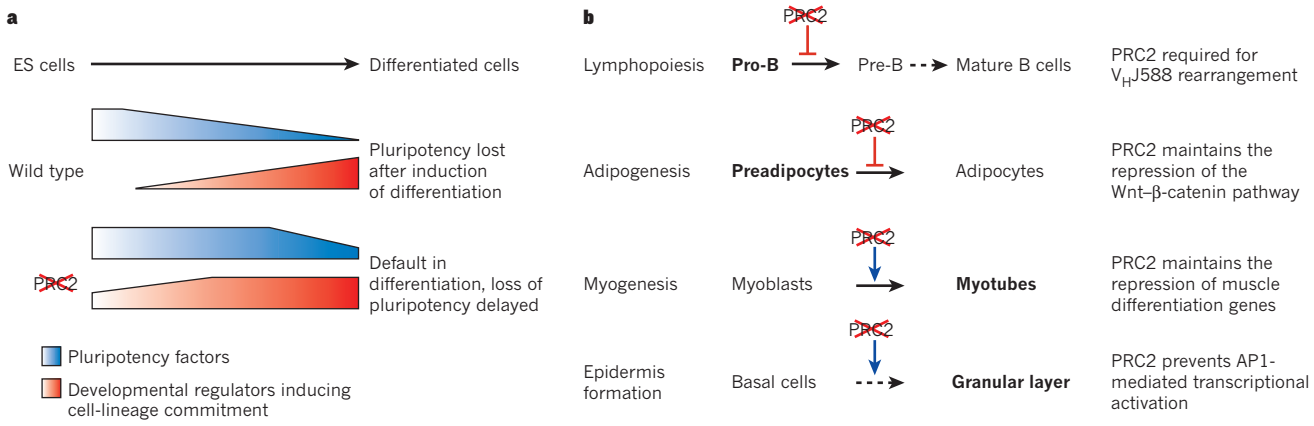


Figure 4 | PRC2-mediated regulation of pluripotency and differentiation. **a**, Comparison of expression levels of pluripotency factors and factors that induce cell commitment during ES-cell differentiation in wild-type and PRC2-impaired ES cells. **b**, The consequences of PRC2 inactivation for cell differentiation. For simplicity, intermediate differentiation steps are not shown in some cases (dashed arrows).

PRC2 and cancer

The expression of PRC2 components is upregulated in various cancers such as melanoma, lymphoma, and breast and prostate cancer. EZH2 has been reported to be a marker of the aggressive stages of prostate and breast malignancies[88,89], and its overexpression promotes neoplastic transformation of normal prostatic cells[90] and hyperplasia in breast epithelium[89,91]. With the exception of the EZH1 homologue, the expression of PRC2 components is regulated by the retinoblastoma protein (pRB)–E2F transcription factors pathway, and is therefore associated with cell proliferation[14,92]. In addition, several microRNAs control EZH2 expression, the deregulation of which could contribute to EZH2 overexpression in cancer. Deletion of PRC2 components in somatic cells led to a marked reduction in cell proliferation[88,92], an effect that was linked to the PRC2-dependent regulation of the *Ink4/Arf* locus (also known as the *Cdkn2a* locus), which encodes the tumour-suppressor proteins $p16^{Ink4a}$, $p19^{Arf}$ and $p15^{Ink4b}$ (refs 87, 93). Given these findings, *EZH2* was proposed to function as an oncogene[92]. Recurrent somatic mutations that interfere with EZH2 enzymatic activity occur in subtypes of lymphoma[94] and myeloid disorder[95,96], but further studies are required to determine the consequences of these deletions for PRC2 activity. Despite the above-mentioned results, EZH2 inactivation does not inhibit cell proliferation in all model cell lines for prostate cancer[97].

To understand the role of PRC2 in tumour progression, it may be more beneficial to determine whether PRC2 is required for the de-differentiation of somatic cells or for the epithelial to mesenchymal transition, rather than modulating EZH2 levels to gauge its function as a tumour-suppressor protein versus an oncoprotein in a defined cell context. Indeed, the apparent outcome in the latter case is probably dependent on the genetic and epigenetic alterations that initiate cellular transformation. Notably, PRC2 seems to be required for the acquisition of pluripotency, as $Eed^{-/-}$ and $Suz12^{-/-}$ ES cells fail to induce the reprogramming of B cells in a heterokaryon assay[98]. If similar mechanisms operate during the reprogramming of somatic cells and during tumour progression, we would expect EZH2 inhibition to be a good approach towards preventing the transition to advanced stages of cancer. Yet, if the carcinogenic process initiates from cancer stem cells, it will be crucial to attain a better understanding of how PRC2 modulates proliferation and, in particular, why PRC2 deletion inhibits the proliferation of some somatic cells but not ES cells.

Perspective

The progress made in understanding the role of PcG proteins, and especially PRC2, has underscored their versatility. Not only is PRC2 involved in the regulation of a broad array of biological processes, but it also establishes regulatory cues that are stable and propagated throughout development. These cues can be subject to adjustment at every step of differentiation or in response to external stimuli. With such a pivotal role in maintaining the repression of different sets of genes depending on the cell type and the developmental stage, PRC2 must be targeted to chromatin by a coordinated and intricate process, the steps of which may entail specific DNA sequence(s), ncRNAs and the chromatin structure associated with its target genes. However, this model relies on hypotheses that require validation. Such validation entails clarification of how ncRNA can recognize defined genomic locations and the exact mechanism by which JARID2 or PCL proteins contribute to PRC2 recruitment.

Although it has now been clearly established that several components of PRC2 are misregulated in disease, their involvement has not been well defined yet. Mouse models that allow genetic manipulations, in conjunction with direct comparisons at the genome-wide level of normal versus pathogenic tissues in defined genetic backgrounds, may provide solid resources for pinpointing the parameters of PRC2-dictated processes. ■

1. Lewis, P. Pc: Polycomb. *Drosoph. Inf. Ser.* **21,** 69 (1949).
2. Lewis, E. B. A gene complex controlling segmentation in *Drosophila*. *Nature* **276,** 565–570 (1978).
3. Schuettengruber, B. & Cavalli, G. Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice. *Development* **136,** 3531–3542 (2009).
4. Klymenko, T. *et al.* A Polycomb group protein complex with sequence-specific DNA-binding and selective methyl-lysine-binding activities. *Genes Dev.* **20,** 1110–1122 (2006).
5. Scheuermann, J. C. *et al.* Histone H2A deubiquitinase activity of the Polycomb repressive complex PR-DUB. *Nature* **465,** 243–247 (2010).
6. Simon, J. A. & Kingston, R. E. Mechanisms of polycomb gene silencing: knowns and unknowns. *Nature Rev. Mol. Cell Biol.* **10,** 697–708 (2009).
7. Eskeland, R. *et al.* Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Mol. Cell* **38,** 452–464 (2010).
8. Ku, M. *et al.* Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet.* **4,** e1000242 (2008).
9. Sing, A. *et al.* A vertebrate Polycomb response element governs segmentation of the posterior hindbrain. *Cell* **138,** 885–897 (2009).
10. Schoeftner, S. *et al.* Recruitment of PRC1 function at the initiation of X inactivation independent of PRC2 and silencing. *EMBO J.* **25,** 3110–3122 (2006).
11. Whitcomb, S. J., Basu, A., Allis, C. D. & Bernstein, E. Polycomb Group proteins: an evolutionary perspective. *Trends Genet.* **23,** 494–502 (2007).
12. Shaver, S., Casas-Mollano, J. A., Cerny, R. L. & Cerutti, H. Origin of the polycomb repressive complex 2 and gene silencing by an E(z) homolog in the unicellular alga *Chlamydomonas*. *Epigenetics* **5,** 301–302 (2010).
13. Ohno, K., McCabe, D., Czermin, B., Imhof, A. & Pirrotta, V. ESC, ESCL and their roles in Polycomb Group mechanisms. *Mech. Dev.* **125,** 527–541 (2008).
14. Margueron, R. *et al.* Ezh1 and Ezh2 maintain repressive chromatin through different mechanisms. *Mol. Cell* **32,** 503–518 (2008).
15. Hennig, L. & Derkacheva, M. Diversity of Polycomb group complexes in plants: same rules, different players? *Trends Genet.* **25,** 414–423 (2009).
16. Shen, X. *et al.* EZH1 mediates methylation on histone H3 lysine 27 and complements EZH2 in maintaining stem cell identity and executing pluripotency. *Mol. Cell* **32,** 491–502 (2008).
17. Cao, R. & Zhang, Y. SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED–EZH2 complex. *Mol. Cell* **15,** 57–67 (2004).
18. Kim, H., Kang, K. & Kim, J. AEBP2 as a potential targeting protein for Polycomb Repression Complex PRC2. *Nucleic Acids Res.* **37,** 2940–2950 (2009).
19. Wang, S., Robertson, G. P. & Zhu, J. A novel human homologue of *Drosophila* polycomblike gene is up-regulated in multiple cancers. *Gene* **343,** 69–78 (2004).
20. Nekrasov, M. *et al.* Pcl-PRC2 is needed to generate high levels of H3-K27 trimethylation at Polycomb target genes. *EMBO J.* **26,** 4078–4088 (2007).
21. Walker, E. *et al.* Polycomb-like 2 associates with PRC2 and regulates transcriptional networks during mouse embryonic stem cell self-renewal and differentiation. *Cell Stem Cell* **6,** 153–166 (2010).
22. Li, G. *et al.* Jarid2 and PRC2, partners in regulating gene expression. *Genes Dev.* **24,** 368–380 (2010).
23. Sarma, K., Margueron, R., Ivanov, A., Pirrotta, V. & Reinberg, D. Ezh2 requires PHF1 to efficiently catalyze H3 lysine 27 trimethylation *in vivo*. *Mol. Cell. Biol.* **28,** 2718–2731 (2008).
24. Savla, U., Benes, J., Zhang, J. & Jones, R. S. Recruitment of *Drosophila* Polycomb-group proteins by Polycomblike, a component of a novel protein complex in larvae. *Development* **135,** 813–817 (2008).
25. Jung, J., Mysliwiec, M. R. & Lee, Y. Roles of JUMONJI in mouse embryonic development. *Dev. Dyn.* **232,** 21–32 (2005).
26. Peng, J. *et al.* Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. *Cell* **139,** 1290–1302 (2009).
27. Shen, X. *et al.* Jumonji modulates polycomb activity and self-renewal versus differentiation of stem cells. *Cell* **139,** 1303–1314 (2009).
28. Pasini, D. *et al.* JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. *Nature* **464,** 306–310 (2010).
29. Landeira, D. *et al.* Jarid2 is a PRC2 component in embryonic stem cells required for multi-lineage differentiation and recruitment of PRC1 and RNA Polymerase II to developmental regulators. *Nature Cell Biol.* **12,** 618–624 (2010).
30. Zee, B. M. *et al. In vivo* residue-specific histone methylation dynamics. *J. Biol. Chem.* **285,** 3341–3350 (2010).
31. Peters, A. H. *et al.* Partitioning and plasticity of repressive histone methylation states in mammalian chromatin. *Mol. Cell* **12,** 1577–1589 (2003).
32. Tie, F. *et al.* CBP-mediated acetylation of histone H3 lysine 27 antagonizes *Drosophila* Polycomb silencing. *Development* **136,** 3131–3141 (2009).
33. Trojer, P. & Reinberg, D. Facultative heterochromatin: is there a distinctive molecular signature? *Mol. Cell* **28,** 1–13 (2007).
34. Cui, K. *et al.* Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell* **4,** 80–93 (2009).
35. Jacob, Y. *et al.* ATXR5 and ATXR6 are H3K27 monomethyltransferases required for chromatin structure and gene silencing. *Nature Struct. Mol. Biol.* **16,** 763–768 (2009).
36. Pasini, D., Bracken, A. P., Hansen, J. B., Capillo, M. & Helin, K. The polycomb group protein Suz12 is required for embryonic stem cell differentiation. *Mol. Cell. Biol.* **27,** 3769–3779 (2007).
37. Swigut, T. & Wysocka, J. H3K27 demethylases, at long last. *Cell* **131,** 29–32 (2007).
38. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837 (2007).

39. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).
40. Stock, J. K. *et al.* Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. *Nature Cell Biol.* **9,** 1428–1435 (2007).
41. Zhao, X. D. *et al.* Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell* **1,** 286–298 (2007).
42. Kanhere, A. *et al.* Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. *Mol. Cell* **38,** 675–688 (2010).
43. Mohn, F. *et al.* Lineage-specific polycomb targets and *de novo* DNA methylation define restriction and potential of neuronal progenitors. *Mol. Cell* **30,** 755–766 (2008).
This paper compares genome enrichment of H3K27me3, H3K4me3 and DNA methylation in ES cells with that in terminally differentiated neurons, demonstrating the plasticity of these marks.
44. Lee, T. I. *et al.* Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125,** 301–313 (2006).
45. Boyer, L. A. *et al.* Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441,** 349–353 (2006).
46. Bracken, A. P., Dietrich, N., Pasini, D., Hansen, K. H. & Helin, K. Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. *Genes Dev.* **20,** 1123–1136 (2006).
47. Squazzo, S. L. *et al.* Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome Res.* **16,** 890–900 (2006).
48. Schwartz, Y. B. *et al.* Genome-wide analysis of Polycomb targets in *Drosophila melanogaster*. *Nature Genet.* **38,** 700–705 (2006).
49. Tolhuis, B. *et al.* Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in *Drosophila melanogaster*. *Nature Genet.* **38,** 694–699 (2006).
50. Hawkins, R. D. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6,** 479–491 (2010).
51. Pan, G. *et al.* Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell* **1,** 299–312 (2007).
52. Rosenfeld, J. A. *et al.* Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics* **10,** 143 (2009).
53. Leeb, M. *et al.* Polycomb complexes act redundantly to repress genomic repeats and genes. *Genes Dev.* **24,** 265–276 (2010).
54. Margueron, R. & Reinberg, D. Chromatin structure and the inheritance of epigenetic information. *Nature Rev. Genet.* **11,** 285–296 (2010).
55. Mattout, A. & Meshorer, E. Chromatin plasticity and genome organization in pluripotent embryonic stem cells. *Curr. Opin. Cell Biol.* **22,** 334–341 (2010).
56. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–326 (2006).
57. Vastenhouw, N. L. *et al.* Chromatin signature of embryonic pluripotency is established during genome activation. *Nature* **464,** 922–926 (2010).
58. Schuettengruber, B. *et al.* Functional anatomy of polycomb and trithorax chromatin landscapes in *Drosophila* embryos. *PLoS Biol.* **7,** e13 (2009).
59. Creyghton, M. P. *et al.* H2AZ is enriched at polycomb complex target genes in ES cells and is necessary for lineage commitment. *Cell* **135,** 649–661 (2008).
This paper reports co-localization of the histone variant H2Az with PRC2 in undifferentiated ES cells, illustrating changes in chromatin structure while cells differentiate.
60. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454,** 766–770 (2008).
61. Brunner, A. L. *et al.* Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res.* **19,** 1044–1056 (2009).
62. Zemach, A., McDaniel, I. E., Silva, P. & Zilberman, D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328,** 916–919 (2010).
63. Woo, C. J., Kharchenko, P. V., Daheron, L., Park, P. J. & Kingston, R. E. A region of the human HOXD cluster that confers polycomb-group responsiveness. *Cell* **140,** 99–110 (2010).
64. Wilkinson, F. H., Park, K. & Atchison, M. L. Polycomb recruitment to DNA *in vivo* by the YY1 REPO domain. *Proc. Natl Acad. Sci. USA* **103,** 19296–19301 (2006).
65. Xi, H. *et al.* Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. *Genome Res.* **17,** 798–806 (2007).
66. Plath, K. *et al.* Role of histone H3 lysine 27 methylation in X inactivation. *Science* **300,** 131–135 (2003).
67. Maenner, S. *et al.* 2-D structure of the A region of Xist RNA and its implication for PRC2 association. *PLoS Biol.* **8,** e1000276 (2010).
68. Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322,** 750–756 (2008).
69. Kohlmaier, A. *et al.* A chromosomal memory triggered by Xist regulates histone methylation in X inactivation. *PLoS Biol.* **2,** E171 (2004).
70. Pandey, R. R. *et al.* Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol. Cell* **32,** 232–246 (2008).
71. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129,** 1311–1323 (2007).
72. Tsai, M. C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329,** 689–693 (2010).
This paper shows a widespread role for HOTAIR ncRNA in the regulation of PRC2 gene targeting, and suggests that HOTAIR bridges PRC2 and LSD1.
73. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl Acad. Sci. USA* **106,** 11667–11672 (2009).
74. Song, J. J., Garlick, J. D. & Kingston, R. E. Structural basis of histone H4 recognition by p55. *Genes Dev.* **22,** 1313–1318 (2008).
75. Margueron, R. *et al.* Role of the polycomb protein EED in the propagation of repressive histone marks. *Nature* **461,** 762–767 (2009).
This paper reports that PRC2 function is regulated by the mark it deposits, thus providing a potential mechanism for the spreading of this mark.
76. Chen, S. *et al.* Cyclin-dependent kinases regulate epigenetic gene silencing through phosphorylation of EZH2. *Nature Cell Biol.* **12,** 1108–1114 (2010).
77. Kaneko, S. *et al.* Phosphorylation of the PRC2 component Ezh2 is cell cycle-regulated and upregulates its binding to HOTAIR ncRNA. *Genes Dev.* **24,** 2615–2620 (2010).
78. Yap, K. L. *et al.* Molecular interplay of the noncoding RNA *ANRIL* and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of *INK4a*. *Mol. Cell* **38,** 662–674 (2010).
In this paper, the authors suggested that ncRNA and H3K27me3 can work together to contribute to PRC1 recruitment.
79. Francis, N. J., Follmer, N. E., Simon, M. D., Aghia, G. & Butler, J. D. Polycomb proteins remain bound to chromatin and DNA during DNA replication *in vitro*. *Cell* **137,** 110–122 (2009).
80. Hansen, K. H. *et al.* A model for transmission of the H3K27me3 epigenetic mark. *Nature Cell Biol.* **10,** 1291–1300 (2008).
81. Chamberlain, S. J., Yee, D. & Magnuson, T. Polycomb repressive complex 2 is dispensable for maintenance of embryonic stem cell pluripotency. *Stem Cells* **26,** 1496–1505 (2008).
82. Yuzyuk, T., Fakhouri, T. H., Kiefer, J. & Mango, S. E. The polycomb complex protein *mes-2/E(z)* promotes the transition from developmental plasticity to differentiation in *C. elegans* embryos. *Dev. Cell* **16,** 699–710 (2009).
83. Endoh, M. *et al.* Polycomb group proteins Ring1A/B are functionally linked to the core transcriptional regulatory circuitry to maintain ES cell identity. *Development* **135,** 1513–1524 (2008).
84. Su, I. H. *et al.* Ezh2 controls B cell development through histone H3 methylation and *Igh* rearrangement. *Nature Immunol.* **4,** 124–131 (2003).
85. Wang, L., Jin, Q., Lee, J. E., Su, I. H. & Ge, K. Histone H3K27 methyltransferase Ezh2 represses *Wnt* genes to facilitate adipogenesis. *Proc. Natl Acad. Sci. USA* **107,** 7317–7322 (2010).
86. Caretti, G., Di Padova, M., Micales, B., Lyons, G. E. & Sartorelli, V. The Polycomb Ezh2 methyltransferase regulates muscle gene expression and skeletal muscle differentiation. *Genes Dev.* **18,** 2627–2638 (2004).
87. Ezhkova, E. *et al.* Ezh2 orchestrates gene expression for the stepwise differentiation of tissue-specific stem cells. *Cell* **136,** 1122–1135 (2009).
88. Varambally, S. *et al.* The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* **419,** 624–629 (2002).
89. Kleer, C. G. *et al.* EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proc. Natl Acad. Sci. USA* **100,** 11606–11611 (2003).
90. Karanikolas, B. D., Figueiredo, M. L. & Wu, L. Polycomb group protein enhancer of zeste 2 is an oncogene that promotes the neoplastic transformation of a benign prostatic epithelial cell line. *Mol. Cancer Res.* **7,** 1456–1465 (2009).
91. Li, X. *et al.* Targeted overexpression of EZH2 in the mammary gland disrupts ductal morphogenesis and causes epithelial hyperplasia. *Am. J. Pathol.* **175,** 1246–1254 (2009).
92. Bracken, A. P. *et al. EZH2* is downstream of the pRB-E2F pathway, essential for proliferation and amplified in cancer. *EMBO J.* **22,** 5323–5335 (2003).
93. Bracken, A. P. *et al.* The Polycomb group proteins bind throughout the *INK4A-ARF* locus and are disassociated in senescent cells. *Genes Dev.* **21,** 525–530 (2007).
94. Morin, R. D. *et al.* Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nature Genet.* **42,** 181–185 (2010).
This study is the first report that somatic mutations resulting in the inactivation of PRC2 are found in diseases.
95. Ernst, T. *et al.* Inactivating mutations of the histone methyltransferase gene *EZH2* in myeloid disorders. *Nature Genet.* **42,** 722–726 (2010).
96. Nikoloski, G. *et al.* Somatic mutations of the histone methyltransferase gene *EZH2* in myelodysplastic syndromes. *Nature Genet.* **42,** 665–667 (2010).
97. Karanikolas, B. D., Figueiredo, M. L. & Wu, L. Comprehensive evaluation of the role of EZH2 in the growth, invasion, and aggression of a panel of prostate cancer cell lines. *Prostate* **70,** 675–688 (2010).
98. Pereira, C. F., Piccolo, F. M., Tsubouchi, T., Sauer, S. & Ryan, N. ESCs require PRC2 to direct the successful reprogramming of differentiated cells toward pluripotency. *Cell Stem Cell* **6,** 547–556 (2010).

Acknowledgements We are grateful to L. Vales, E. Heard and R. Bonasio for crucial reading of this manuscript and active discussions. We apologize to authors whose studies could not be cited owing to space limitations. Work in the laboratory of R.M. is supported by the Institut Nationale du Cancer and Fondation pour la Recherche Medicale. Work in the laboratory of D.R. is funded by the US National Institutes of Health (grants RO1GM064844 and 4R37GM037120) and the Howard Hughes Medical Institute.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to D.R. (danny.reinberg@nyumc.org).

PERSPECTIVE

doi:10.1038/nature09659

Systemic risk in banking ecosystems

Andrew G. Haldane[1] & Robert M. May[2]

In the run-up to the recent financial crisis, an increasingly elaborate set of financial instruments emerged, intended to optimize returns to individual institutions with seemingly minimal risk. Essentially no attention was given to their possible effects on the stability of the system as a whole. Drawing analogies with the dynamics of ecological food webs and with networks within which infectious diseases spread, we explore the interplay between complexity and stability in deliberately simplified models of financial networks. We suggest some policy lessons that can be drawn from such models, with the explicit aim of minimizing systemic risk.

In the 1960s, the notion of the 'balance of nature' played a significant part as ecologists sought a conceptual foundation for their subject. In particular, Evelyn Hutchinson[1], following Elton[2], suggested that "oscillations observed in arctic and boreal fauna may be due in part to the communities not being sufficiently complex to damp out oscillations". He went on to state, based on a misunderstanding of MacArthur's[3] paper, that there was now a "formal proof of the increase in stability of a community as the number of links in its food web increases".

To the direct contrary, however, a closer examination of model ecosystems showed that a random assembly of N species, each of which had feedback mechanisms that would ensure the population's stability were it alone, showed a sharp transition from overall stability to instability as the number and strength of interactions among species increased. More explicitly, for $N \gg 1$ this transition occurs once $m\alpha^2 > 1$, where m is the average number of links per species, and $(\pm)\ \alpha$ their average strength[4].

In ecology this has, since the 1970s, prompted a search for special food-web structures that may help reconcile complexity with persistence or stability[5–8]. Along these lines there is, for example, tentative evidence for modularity[9] (particularly in plant–pollinator associations, where linkages tend to be overdispersed or disassociative), and more generally for nested hierarchies in food webs[10]. The fact that some features of the network structure of interactions (such as predator/prey ratios) inferred from the Burgess Shale communities are similar to those in present day ones[11] reinforces hopes that this is a meaningful area of research.

In the wake of the global financial crisis that began in 2007, there is increasing recognition of the need to address risk at the systemic level, as distinct from focusing on individual banks[12,13]. This quest to understand the network dynamics of what might be called 'financial ecosystems' has interesting parallels with ecology in the 1970s. Implicit in much economic thinking in general, and financial mathematics in particular, is the notion of a 'general equilibrium'. Elements of this belief underpin, for example, the pricing of complex derivatives. But, as shown below, deeper analysis of such systems reveals explicit analogies with the concept that too much complexity implies instability, which was found earlier in model ecosystems.

There are, of course, major differences between ecosystems and financial systems. For one thing, today's ecosystems are the winnowed survivors of long-lasting evolutionary processes, whereas the evolution of financial systems is a relatively recent phenomenon[14]. Nor have selective pressures been entirely dispassionate, with the hand of government a constant presence shaping financial structures, especially among institutions deemed "too big to fail"[15]. In financial ecosystems, evolutionary forces have often been survival of the fattest rather than the fittest.

In what follows, we first consider the role of the growth in intrafinancial system claims in generating bank failure and instability, focusing on the problems inherent in prevailing methods of pricing complex derivatives, or arbitrage pricing theory (APT). Second, we sketch various ways in which such an initial bank failure, or 'shock', may propagate to cause cascades of subsequent failure. Third, we outline some tentative policy lessons that might be drawn from these deliberately oversimplified models. Last, we ask how we might reshape the financial system to realize the economic benefits individual banks can deliver, while at the same time paying deliberate and explicit attention to their system-wide stability.

Potential causes of an initial shock

Events external to the banking system, such as recessions, major wars, civil unrest or environmental catastrophes, clearly have the potential to depress the value of a bank's assets so severely that the system fails. Although probably exacerbated by such events, including global imbalances (China as producer and saver, the United States as consumer and debtor), the present crisis seems more akin to self-harm caused by overexuberance within the financial sector itself. Perhaps as much as two-thirds of the spectacular growth in banks' balance sheet over recent decades reflected increasing claims within the financial system, rather than with non-financial agents. One key driver of this explosive intrasystem activity came from the growth in derivative markets.

In 2002, when Warren Buffet first expressed his view that "derivatives are financial weapons of mass destruction"[16], markets—although booming—seemed remarkably stable. Their subsequent growth, illustrated in Fig. 1, has been extraordinary, outpacing the growth in world gross domestic product (GDP) by a factor of three. In some derivatives markets, such as credit default swaps (CDS), growth has outpaced Moore's Law. These developments contributed significantly towards an unprecedented influx of mathematically skilled people (quantitative analysts) into the financial/banking industry. These people produced very sophisticated techniques (including APT), which seemingly allowed you to put a price on future risks, and thus to trade increasingly complex derivative contracts—bundles of assets—with risks apparently decreasing as the bundles grew.

However, recent empirical and theoretical studies have indicated that the trading activity associated with derivatives can have significant effects on markets[17–19]. More specifically, Brock and colleagues[20] have shown that proliferation of hedging instruments can destabilize markets. Building on this, Caccioli and colleagues[21] note that APT makes several conventional assumptions upon which everything else depends: "perfect competition, market liquidity, no-arbitrage and market completeness". Crucially, this adds up to the implicit assumption that trading activity has no feedback on the dynamical behaviour of markets. And indeed, in the APT-fuelled

[1]Bank of England, Threadneedle Street, London EC2R 8AH, UK. [2]Zoology Department, Oxford University, Oxford OX1 3PS, UK.

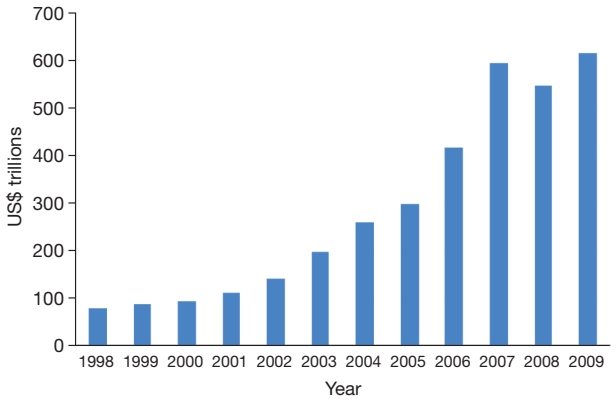


Figure 1 | Notional principal value of outstanding derivative contracts, as recorded at year end. These include foreign exchange, interest rates, equities, commodities and credit derivatives. Data from UK Department for Business, Innovation and Skills, International Monetary Fund and Bank of England calculations.

boom time that preceded the bust, APT seemed to be very successful. In its imaginary world, market failures are caused by regulatory carelessness, resulting in a focus on creating institutional arrangements that seek to guarantee the premises upon which APT is based[22]. To the contrary, Caccioli and colleagues argued[21] that APT is not a 'theory' in the sense habitually used in the sciences, but rather a set of idealized assumptions on which financial engineering is based; that is, APT is part of the problem itself.

Caccioli and colleagues[21] illustrate their point by exploring the dynamical properties of a model that gives a more realistic caricature of markets, going beyond the idealized world of APT to include the effects of individual trades on prices. Prices now depend on the balance between demand and supply. The outcome is that "the road to efficient, arbitrage-free, complete markets can be plagued by singularities which arise upon increasing financial complexity"[21].

Figure 2 illustrates the main results of the analysis by Caccioli and colleagues[21]. Here n is essentially a measure of the proliferation of derivatives or similar financial instruments, and s is the overall average value of the supply of any one such derivative/financial instrument. The parameter ε encodes the risk premium that banks require for trading derivatives[21].

We see from Fig. 2 that if n is less than n^* (here $n^* = 4.14$), the average supply of derivatives, s, is relatively steady and essentially independent of the banks' risk premium (as measured by ε). But as market complexity increases, so that n approaches n^*, there is a sharp singularity at $\varepsilon = 0$.

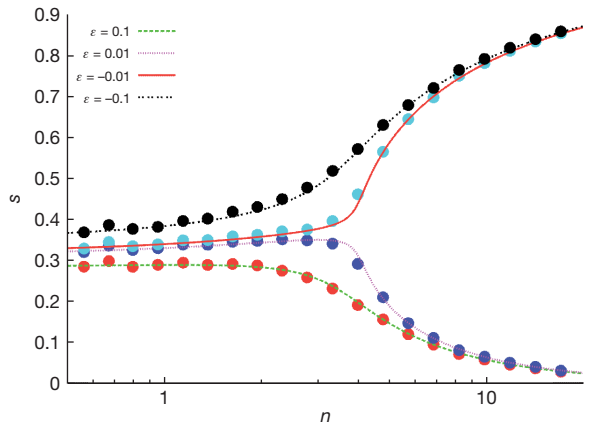


Figure 2 | Discontinuous transition to instability of derivatives as complexity increases. Average supply of any one derivative, s, at competitive equilibrium as a function of the number, n, of different derivatives being traded, for various values of banks' risk premium, ε. Adapted with permission from ref. 21. For fuller discussion, see text.

For $n > n^*$, the average supply increases with increasing n (that is, increasing proliferation of financial instruments) if $\varepsilon > 0$. Conversely, for $\varepsilon < 0$ the supply decreases with increasing complexity once $n > n^*$. It is emphasized[21] that such sensitivity in market behaviour in the neighbourhood of the singularity can easily produce very strong fluctuations—either positive or negative—in the volume of trading in derivative markets.

Note that the consequences of this singularity are not easily intuited from the competitive equilibrium setting. It seems to us that the basic process—in grossly simplified terms—is that once there are enough derivatives to span the space of available states of nature (the net supply of derivatives within the system necessary to meet true hedging demand from non-banks), the market is essentially complete in the sense of the Arrow–Debreu[23] model. Once that happens, gross derivatives positions within the system are essentially unbounded. So long as there is an incentive to supply new instruments—a positive premium to trading—banks will continue to expand gross positions, independent of true hedging demand from non-banks. Such trades are essentially redundant, increasing the dimensionality and complexity of the network at a cost in terms of stability, with no welfare gain because market completeness has already been achieved.

Caccioli and colleagues[21] also examine a measure of market volatility as the risk premium parameter ε varies. If they calculate this quantity under the approximation that the fluctuations in the values of the individual 'supply variables' (s_i; derivatives, etc) are completely uncorrelated, they in effect recover the happy world of APT, with no singularities. This strongly indicates that the highly important singularities in their accurate and self-consistent calculations, with market dynamics included, are associated with the supplies of different derivatives being strongly correlated in this domain, as has found to be the case among derivatives markets in practice.

In summary, Caccioli and colleagues suggest that the idealized assumptions upon which recent financial engineering has been based can give a misleading account of potential instabilities in markets. They also note that these instabilities echo those that can develop in ecosystems as complexity increases[4,24].

Propagation of shocks within financial systems

In ecology's models of food webs, aimed at qualitative understanding of their dynamical response to perturbation, the nodes are simply species, linked to other nodes/species as prey, predator, competitor or mutualist. In epidemiological networks, the nodes are susceptible, infected/infectious or recovered/immune individuals linked by sexual or other contacts. But in a minimally realistic caricature of financial networks—henceforth called banks—the nodes have a more complex structure.

Following Nier and colleagues[25] and Gai and Kapadia[26], we define such a bank/node as schematically illustrated in Fig. 3. In this deliberately oversimplified scheme, a bank's activities are partitioned among four categories. Two represent assets: interbank loans (l_i) and external assets (e_i). The other two represent liabilities: interbank borrowing (b_i) and deposits (d_i). The subscript i labels the specific bank (i = 1, 2, ..., N for a total of N banks). Solvency requires that the difference between a bank's assets and its liabilities (the capital reserve or 'net worth', labelled γ_i in Fig. 3) be positive. That is, $\gamma_I \equiv (e_i + l_i) - (d_i + b_i) \geq 0$.

These banks are now assumed to be interlinked in a random, Erdős–Rényi network, with any one of the N banks connected to any other as lender or borrower, or possibly both, each with probability p. A bank's average number of incoming/borrowing or outgoing/lending links is then $z = p(N - 1)$.

Various further assumptions are now made to carry these Bank of England/Federal Reserve Bank of New York models to the point where the knock-on effects of a single bank failure can be explored in numerical simulations. Much of the essential findings of such studies can be captured, and made more transparent, by a 'mean-field' approximation in which each bank has exactly average behaviour[27]. This means all banks are the same size (rescaled to 1), every bank is linked to exactly z others, all loans

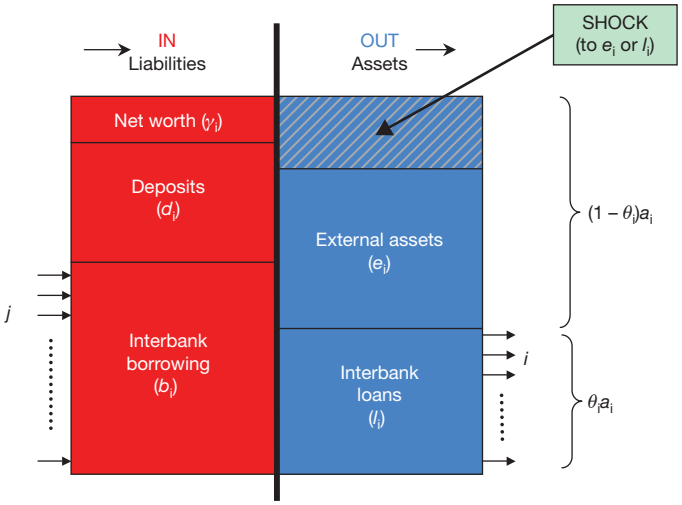


Figure 3 | Schematic model for a node in the interbank network. Adapted with permission from ref. 25.

have the same magnitude, w, as do the capital reserves, γ, and the ratios of loans to total assets, θ.

As illustrated in Fig. 3, all these models study the consequences of a shock that initially hits a single bank, wiping out a fraction, f, of its external assets. If the magnitude of this shock exceeds the capital reserve, $f(1-\theta) > \gamma$, the bank fails. This is a deliberate oversimplification, aimed at a clearer understanding of how an initial failure can propagate shocks throughout the system.

The most direct effect of such a failure is that its z creditor banks will lose part or all of their loans. If such losses exceed γ, these banks in turn will fail, propagating a third phase of shocks to those remaining, and so on. Note, however, that a failing bank's losses are in effect divided among its z creditors, so that each subsequent phase of loan-driven shocks is attenuated, approximately by a factor z.

Figure 4 illustrates one of the tentative messages emerging from this toy model, showing regimes of failure in terms of the critical parameter γ (capital reserves relative to bank size) and θ (interbank activity as a fraction of total assets). Within the unhatched triangle (0, 1, f), the initially shocked bank fails; in the blue triangle (0, 1, A) a second tranche

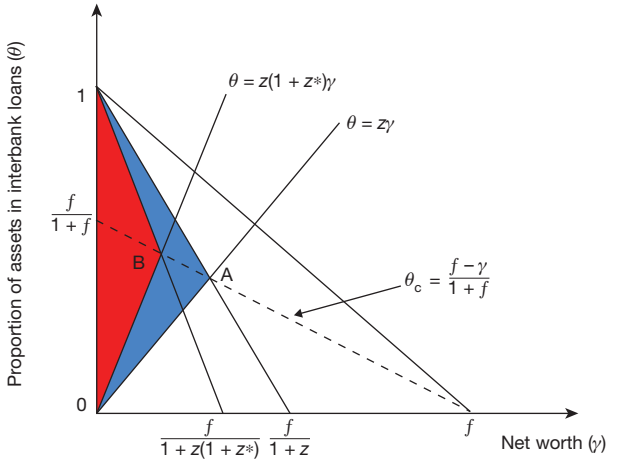


Figure 4 | Domains of interbank lending. Domains are expressed as a fraction of total assets, θ, and capital reserves or net worth, γ, which result in the propagation of interbank loan shocks. The triangle (1, 0, f) defines the region where loss of a fraction f of a bank's external assets will cause it to fail. The blue triangle (1, 0, A) depicts the region in which creditors of the initially failing bank will receive phase II shocks which cause them also to fail, and the red area (1, 0, B) shows the region in which phase III shocks cause failure. Adapted with permission from ref. 27.

of z banks go down; in the red triangle (0, 1, B) there is a third phase of roughly z^2 failures; and so on. Note that, when (as above) the initial shock is to external assets, the system's fragility is maximized (failures for relatively large values of γ) by θ having values intermediate between 0 and 1, which in some ways very roughly corresponds to banks substantially engaged in both retail and investment (high-street and casino) activity. As seen earlier and in Fig. 4, an increase in the system's connectivity, z, causes the coloured region of instability to shrink; high connectivity distributes, and thereby attenuates risk. On the other hand, when later-phase failures do occur, they will then involve more banks.

A second, and almost surely more important, source of shock propagation arises from losses in the value of a bank's external assets, caused by a generalized fall in market prices, a rise in expected defaults or a failing bank's 'fire sale' actions. Such market liquidity shocks are conventionally and sensibly represented by discount factors that, for a given asset class, are proportional to the number of failing banks holding the asset. This may be generalized to distinguish between strong liquidity shocks, associated with discounting specific asset classes, and weak liquidity shocks, resulting from the expectation of further defaults or a more general loss of confidence[27]. In all cases and in sharp contrast to the attenuation in interbank loan shocks, liquidity shocks amplify as more banks fail. Thus, relatively small initial liquidity shocks have the potential to make strong contributions to systemic risk.

A third mechanism of shock propagation, which has been a marked—and in many peoples' opinion the most important—feature of the recent crisis has been the diminished availability of interbank loans, or in the jargon of the trade, 'funding liquidity shocks'. This has often taken the form of liquidity hoarding in interbank funding markets. Gai and Kapadia[28] have recently shown how such liquidity hoarding can cascade through a banking network, with severe consequences. As one bank calls in or shortens the term of its interbank loans, affected banks tend in turn to do the same. The result is a liquidity-hoarding shock that is not subject to the attenuation characteristic of interbank default shocks.

All three propagation mechanisms can be drawn together within the framework defined by Fig. 3 (see also N. Arinaminpathy, S. Kapadia and R.M.M., manuscript in preparation). The model can also be generalized to treat banks of varying size, including the extreme but realistic case of a few very large all-purpose banks, each connected to many smaller banks; interconnectivity within real banking networks is far from random[29–31], with long-tailed degree distributions. It also seems that these networks tend to be disassociative rather than proportionately connected: that is, big banks are disproportionately linked to smaller ones, and conversely. Such a 'wiring up' of a network is known, unfortunately, to maximize the number of individuals infected by an agent that is transmitted by interpersonal contact[32]. On the other hand, such disassociative structures are likely to support a larger number of coexisting banks (another link between ecology and banking[33]), and can make the network more robust to random losses[9,34].

Some of this work, particularly that on liquidity shocks, echoes an important insight from pervious work[35,36] (N. Beale and colleagues, manuscript in preparation). This is that excessive homogeneity within a financial system—all the banks doing the same thing—can minimize risk for each individual bank, but maximize the probability of the entire system collapsing. A very simple toy model illustrates this. Suppose you have N banks and N distinct, uncorrelated asset classes, each of which has some very small probability, ε, of having its value decline to the extent that a bank holding solely that asset would fail. At the inhomogeneous extreme, assume each bank holds the entirety of one of the N assets: the probability for any one bank to fail is now ε, whereas that for the system is a vastly smaller ε^N. At the opposite, homogeneous extreme, assume all banks are identical, each holding $1/N$ of every one of the N assets: the probability for any one bank to fail can now be calculated as $N^N \varepsilon^N/N!$, and this is obviously also the probability for all N of these banks to fail. This homogeneous, 'herding behaviour' limit clearly makes each individual bank safer, but the systemic risk is much larger. More realistic versions of this scenario consistently show the same unhappy

conclusion. Tentative evidence comes from the fact that the world's five largest banks have shown increasing concentrations of assets over the last ten years, in contrast to the top five hedge funds, whose less concentrated systems can give greater scope for diversity. The former are in trouble, the latter much less so.

Implications for public policy

All the studies described earlier involve numerical simulations, but many combine such work with analytic results of the kind exemplified by Fig. 4. Such analysis of the dynamics of deliberately oversimplified models of financial ecosystems carries potentially far-reaching implications for the design and implementation of public policy. These implications include the following.

Setting regulatory capital/liquidity ratios

The cornerstone of the current international regulatory agenda is the setting of higher requirements for banks' capital and liquid assets. The traditional rationale for such requirements is that they reduce idiosyncratic risks to the balance sheets of individual banks. An alternative and more far-reaching interpretation is that they are a means of strengthening the financial system as a whole by limiting the potential for network spillovers. With this wider objective, prudential regulation is following in the footsteps of ecology, which has increasingly drawn on a system-wide perspective when promoting and managing ecosystem resilience.

The systemic rationale for financial regulatory intervention is well illustrated by the dynamic models outlined earlier. Consider banks' buffers of capital or net worth (γ). These capital ratios have been in secular decline in relation to banks' total assets for at least the past 150 years in the United Kingdom and United States[37]. Reversing these trends by setting higher required capital ratios strengthens the absorptive capacity of each of the nodes in the financial network in response to external shocks. As importantly, however, it also lessens the risk of idiosyncratic defaults cascading around the system, as illustrated in Fig. 4.

Broadly, the same arguments apply in the setting of regulatory requirements on banks' liquid assets. These liquidity ratios have also been in secular decline in the United Kingdom and United States, for at least the past half century. Typically, liquidity requirements are specified as a minimum ratio of banks' liquid assets to their short-term liabilities. This liquidity ratio can be seen as a means of short-circuiting the potential for systemic liquidity spillovers arising from fire sales on the asset side of the balance sheet (liquidity shocks) or liquidity hoarding on the liabilities side (liquidity-hoarding shocks). In particular, holdings of liquid assets reduce the potential for market liquidity risk to propagate around the system, while limits on short-term liabilities reduce the spread of funding liquidity risk around the system.

Setting systemic regulatory requirements

Looking at financial risk through a network lens indicates a fundamentally different rationale for prudential regulation. It also indicates a quite different calibration of such regulation. Prudential regulation has become increasingly risk-based with the advent of first Basel I and latterly Basel II. But the risk in question to which regulation was then calibrated has tended to be institution-specific rather than systemic risk.

To give an example, as conventionally calibrated, capital regulation seeks to equalize failure probabilities across individual institutions to a given tolerance threshold—such as a 0.1% probability of failure. Approaching this problem from a system-wide angle indicates a rather different calibration. Instead, the objective would be to set firms' capital requirements to equalize the marginal cost to the system as a whole of their failure. In other words, regulatory requirements would be set higher for those banks bringing greatest risk to the system; for example, because of their size or connectivity.

Although new in the context of banking, the essential insight here is an old one in the study of epidemiological networks. Anderson and May[38] established the theoretical case for focusing preventative action on 'super-spreaders' within the network to limit the potential for system-wide spread. Although initially applied in the study of contagious diseases, such as HIV/AIDs, this same insight has since been applied in managing the dynamics of the world wide web, power grids and biological ecosystems[8,39].

If anything, this same logic applies with even greater force in banking. There has been a spectacular rise in the size and concentration of the financial system over the past two decades, with the rapid emergence of 'super-spreader institutions' too big, connected or important to fail (Fig. 5). The collateral damage, to both the real economy and financial system, following the failure of Lehman Brothers in October 2008 is testimony to the force of such super-spreader dynamics. Protecting the financial system from future such events would require the key super-spreader nodes to run with higher—potentially much higher—buffers of capital and liquid assets, which are then proportional to the system-wide risk they contribute.

A second source of system-wide risk, in addition to super-spreader failures, arises from aggregate external events, such as booms and busts in the real economy. Indeed, historically this has been the largest single source of banking problems. If regulation could be operated countercyclically, with buffers rising in booms and falling in recessions, this would lessen systemic risk from this particular source. Why? Because increasing insurance in a boom would increase system-wide resilience against the subsequent bust, as well as providing an incentive for banks to curb risk-taking during the boom. Operating regulation in this way would be a new departure for prudential policy—so-called macroprudential policy—but a potentially important one[40,41] from a systemic risk perspective.

Netting and clearing derivatives

The rapid growth in the size and complexity of the derivatives market contributed importantly to the destabilizing dynamics of the system under stress during the recent financial crisis. This begs questions about the underlying structure and dimensionality of the derivatives market. One means of simplifying the complex web of interactions between banks in derivatives markets is to centralize the trading and clearing of these instruments. For example, central counterparties interpose themselves between every bilateral transaction, thereby replacing a cat's-cradle of financial network interactions with a single hub-and-spokes configuration. Provided the central counterparty is extremely robust—to prevent it becoming a super-spreader itself—the upshot is

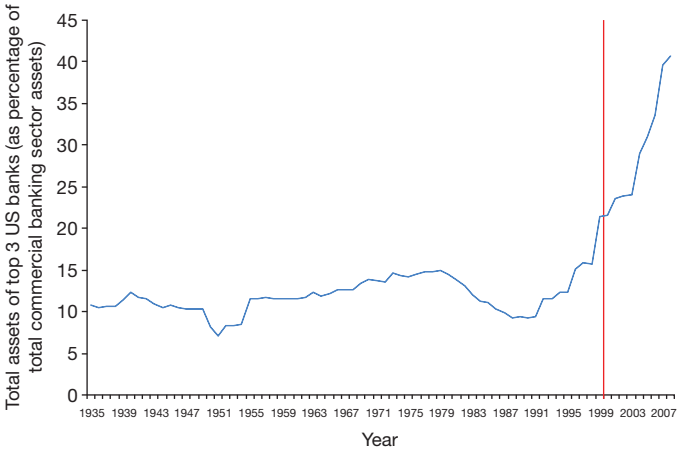


Figure 5 | Recent rise in the size and concentration of the United States financial system. This figure illustrates the marked increase in asset concentration within the United States banking system since the Glass–Steagal restrictions were revoked in 1999. Red line represents the Gramm–Leach–Bliley Act (1999), which revoked Glass–Steagal restrictions. Data include only the insured depository subsidiaries of banks to ensure consistency over time; for example, non-deposit subsidiaries are not included. Data from the Federal Deposit Insurance Corporation.

a less complex and lower-risk financial network. Efforts are underway internationally to extend the scope and reach of central counterparty clearing, in particular to ensure it covers transactions in complex over-the-counter derivative instruments, such as CDS (D. Duffie and H. Zhu, manuscript in preparation).

In parallel, there are international efforts to reduce the dimensionality of derivatives contracts by eliminating redundant trades and through netting. This redundancy might arise either because contracts have been reassigned to participants (but the claim not extinguished) or because there are perfectly offsetting bilateral transactions between two parties that can be netted. For example, the stock of CDS contracts has already been reduced by around $25 trillion since December 2007 as a result of such netting arrangements. Looking forward, there may be more sophisticated multilateral netting algorithms that can be used to reduce further derivatives balances.

Shaping the topology of the financial network

The analytic model outlined earlier demonstrates that the topology of the financial sector's balance sheet has fundamental implications for the state and dynamics of systemic risk. From a public policy perspective, two topological features are key[15].

First, diversity across the financial system. In the run-up to the crisis, and in the pursuit of diversification, banks' balance sheets and risk management systems became increasingly homogenous. For example, banks became increasingly reliant on wholesale funding on the liabilities side of the balance sheet; in structured credit on the assets side of their balance sheet; and managed the resulting risks using the same value-at-risk models. This desire for diversification was individually rational from a risk perspective. But it came at the expense of lower diversity across the system as whole, thereby increasing systemic risk. Homogeneity bred fragility (N. Beale and colleagues, manuscript in preparation).

In regulating the financial system, little effort has as yet been put into assessing the system-wide characteristics of the network, such as the diversity of its aggregate balance sheet and risk management models. Even less effort has been put into providing regulatory incentives to promote diversity of balance sheet structures, business models and risk management systems. In rebuilding and maintaining the financial system, this systemic diversity objective should probably be given much greater prominence by the regulatory community.

Second, modularity within the financial system. The structure of many non-financial networks is explicitly and intentionally modular. This includes the design of personal computers and the world wide web and the management of forests and utility grids. Modular configurations prevent contagion infecting the whole network in the event of nodal failure. By limiting the potential for cascades, modularity protects the systemic resilience of both natural and constructed networks.

The same principles apply in banking. That is why there is an ongoing debate on the merits of splitting banks, either to limit their size (to curtail the strength of cascades following failure) or to limit their activities (to curtail the potential for cross-contamination within firms). The recently proposed Volcker rule in the United States, quarantining risky hedge fund, private equity and proprietary trading activity from other areas of banking business, is one example of modularity in practice. In the United Kingdom, the new government have recently set up a Royal Commission to investigate the case for encouraging modularity and diversity in banking ecosystems, as a means of buttressing systemic resilience.

It took a generation for ecological models to adapt. The same is likely to be true of banking and finance.

1. Hutchinson, G. E. Homage to Santa Rosalia, or why are there so many kinds of animals? *Am. Nat.* **93,** 145–159 (1959).
2. Elton, C. S. *The Ecology of Invasions by Animals and Plants* (Methuen, 1958).
3. MacArthur, R. H. Fluctuations of animal populations, and a measure of community stability. *Ecology* **36,** 533–536 (1955).
4. May, R. M. Will a large complex system be stable? *Nature* **238,** 413–414 (1972).
5. Stouffer, D. B. *et al.* Quantitative patterns in the structure of model and empirical food webs. *Ecology* **86,** 1301–1311 (2005).
6. Pascuale, M. & Dunne, J. A. *Ecological Networks: Linking Structure to Dynamics in Food Webs* (Oxford Univ. Press, 2006).
7. Dunne, J. A. *et al.* Network structure and robustness of marine food webs. *Mar. Ecol. Prog. Ser.* **273,** 291–302 (2004).
8. May, R. M. Network structure and the biology of populations. *Trends Ecol. Evol.* **21,** 394–399 (2006).
9. Bascompte, J. Disentangling the web of life. *Science* **325,** 416–419 (2009).
10. Sugihara, G. & Ye, H. Cooperative network dynamics. *Nature* **458,** 979–980 (2009).
11. Dunne, J. A. *et al.* Compilation and network analyses of Cambrian food webs. *PLoS Biol.* **6,** e102 (2008).
12. Haldane, A. G. Rethinking the financial network. ⟨http://www.bankofengland.co.uk/publications/speeches/2009/speech386.pdf⟩ (2009).
13. Jones, C. Preventing system failure. *Cent. Banking* **21,** 69–75 (2010).
The survey of recent thinking about financial networks, in lay language.
14. Farmer, J. D. Market force, ecology and evolution. *Ind. Corp. Change* **11,** 895–953 (2002).
15. Haldane, A. G. The $100 billion question. ⟨http://www.bankofengland.co.uk/publications/speeches/2010/speech433.pdf⟩ (2010).
16. Buffet, W. E. Chairman's Letter. *Berkshire Hathaway Inc. 2002 Annual Report* 15 (2002).
A prescient and clearly expressed warning of problems inherent in derivatives.
17. Sircar, K. R. & Papanicolau, G. General Black-Scholes models accounting for increased market volatility from hedging strategies. *Appl. Math. Finance* **5,** 45–82 (1998).
18. Avellaneda, M. & Lipkin, M. D. A market-induced mechanism for stock pinning. *Quantit. Finance* **3,** 417–425 (2003).
19. Osler, C. L. Macro lessons from microstructure. *Int. J. Finance Econ.* **11,** 55–80 (2006).
20. Brock, W. A., Hommes, C. H. & Wagner, F. O. O. More hedging instruments may destabilise markets. *J. Econ. Dynam. Cont.* **33,** 1912–1928 (2008).
21. Caccioli, F., Marsili, M. & Vivo, P. Eroding market stability by proliferation of financial instruments. *Eur. Phys. J. B* **71,** 467–479 (2009).
A sophisticated and important analysis of a major flaw in the pricing of derivatives.
22. Pliska, S. R. *Introduction to Mathematical Finance: Discrete Time Models* (Blackwell, 1997).
23. Arrow, K. J. & Debreu, G. Existence of an equilibrium for a competitive economy. *Econometrica* **22,** 265–290 (1954).
24. May, R. M. *Stability and Complexity in Model Ecosystems* (Princeton Univ. Press, 1973).
25. Nier, E., Yang, J., Yorulmazer, T. & Alentorn, A. Network models and financial stability. *J. Econ. Dyn. Control* **31,** 2033–2060 (2007).
26. Gai, P. & Kapadia, S. Contagion in financial networks. *Proc. R. Soc. A* **466,** 2401–2423 (2010).
27. May, R. M. & Arinaminpathy, N. Systemic risk: the dynamics of model banking systems. *J. R. Soc. Interface* **7,** 823–838 (2010).
28. Gai, P. & Kapadia, S. Liquidity hoarding, network externalities, and interbank market collapse. *Proc. R. Soc. A* **466,** 2401–2423 (2010).
29. Schweitzer, F. *et al.* Economic networks: the new challenges. *Science* **325,** 422–425 (2009).
30. May, R. M., Levin, S. A. & Sugihara, G. Complex systems: ecology for bankers. *Nature* **451,** 893–895 (2008).
A brief overview of the 2006 NAS/FRBNY study on systemic risk.
31. Kyriakopoulos, F. *et al.* Network and eigenvalue analysis of financial transaction networks. *Eur. Phys. J. B* **71,** 523–531 (2009).
32. Gupta, S., Anderson, R. M. & May, R. M. Networks of sexual contacts: implications for the pattern of spread of HIV. *AIDS* **3,** 807–818 (1989).
33. Bastolla, U. *et al.* The architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature* **458,** 1018–1020 (2009).
34. Memmott, J. *et al.* Tolerance of pollination networks to species extinctions. *Proc. R. Soc. Lond. B* **271,** 2605–2611 (2004).
35. Battiston, A., Gatti, D. D., Gallegati, M., Greenwald, B. C. & Stiglitz, J. E. Liaisons dangereuses: increasing connectivity, risk sharing and systemic risk. ⟨http://www.nber.org/papers/w15611⟩ (2009).
36. Stiglitz, J. Contagion, liberalization, and the optimal structure of globalization. *J. Global. Develop.* (in the press).
37. Haldane, A. G. Banking on the state. ⟨http://www.bankofengland.co.uk/publications/speeches/2009/speech409.pdf⟩ (2009).
This gives a more full account of the policy issues discussed here.
38. Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Transmission and Control* Ch.12.3 (Oxford Univ. Press, 1991).
39. Barabesi, A.-L. *Linked: The New Science of Networks* (Perseus, 2002).
40. Bank of England. The role of macroprudential policy: a discussion paper. ⟨http://www.bankofengland.co.uk/publications/other/financialstability/roleofmacroprudentialpolicy091121.pdf⟩ (2009).
41. Turner, P. The debate on financial system resilience: macroprudential instruments. ⟨http://www.bankofengland.co.uk/publications/speeches/2009/speech407.pdf⟩ (2009).

Acknowledgements We are indebted to colleagues (particularly S. Kapadia, N. Arinaminpathy and G. Sugihara), who made many helpful comments and constructive criticisms.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to R.M.M. (robert.may@zoo.ox.ac.uk).

ARTICLE

doi:10.1038/nature09650

Genetic variegation of clonal architecture and propagating cells in leukaemia

Kristina Anderson[1], Christoph Lutz[2], Frederik W. van Delft[1], Caroline M. Bateman[1], Yanping Guo[2], Susan M. Colman[1], Helena Kempski[3], Anthony V. Moorman[4], Ian Titley[1], John Swansbury[1], Lyndal Kearney[1], Tariq Enver[2]† & Mel Greaves[1]

Little is known of the genetic architecture of cancer at the subclonal and single-cell level or in the cells responsible for cancer clone maintenance and propagation. Here we have examined this issue in childhood acute lymphoblastic leukaemia in which the *ETV6–RUNX1* gene fusion is an early or initiating genetic lesion followed by a modest number of recurrent or 'driver' copy number alterations. By multiplexing fluorescence *in situ* hybridization probes for these mutations, up to eight genetic abnormalities can be detected in single cells, a genetic signature of subclones identified and a composite picture of subclonal architecture and putative ancestral trees assembled. Subclones in acute lymphoblastic leukaemia have variegated genetics and complex, nonlinear or branching evolutionary histories. Copy number alterations are independently and reiteratively acquired in subclones of individual patients, and in no preferential order. Clonal architecture is dynamic and is subject to change in the lead-up to a diagnosis and in relapse. Leukaemia propagating cells, assayed by serial transplantation in NOD/SCID IL2Rγ^{null} mice, are also genetically variegated, mirroring subclonal patterns, and vary in competitive regenerative capacity *in vivo*. These data have implications for cancer genomics and for the targeted therapy of cancer.

Recent genome-wide scrutiny of cancer cells has revealed extraordinary complexity, with substantial numbers of both potential 'driver' and neutral or 'passenger' mutations per case[1,2]. Informative though these screens are, they probably reflect predominant or composite genetic landscapes that obscure the existence of subclonal heterogeneity of disease[3]. Intraclonal genetic diversity is a common feature of cancer[4] and is probably, from a Darwinian, natural selection perspective, the essential substrate for clonal evolution, disease progression, relapse or metastasis. Subclonal genetic complexity might also be an important consideration for therapeutic targeting. Furthermore, if a subset of 'stem-like' cancer cells, or, as we refer to, propagating cells, are the basis of sustained clonal expansion and disease progression[5] then, in principle, they should be genetically diverse if selection and passage through evolutionary bottlenecks is to occur.

Identifying intraclonal genetic architecture requires genetic scrutiny of single cells or clonal foci, and there are limited examples of this so far[6]; nevertheless, they testify to the existence of significant heterogeneity. The genetic diversity of cancer propagating cells is, as yet, unexplored. We elected to address this issue in lymphoblastic leukaemia. The substantial advantage of this cancer, in addition to its amenability to single-cell analysis, is that it is minimally deranged or unstable, genetically, and the broad, temporal sequence of genetic events is known. For the B-cell precursor subset of childhood acute lymphoblastic leukaemia (ALL) with *ETV6–RUNX1* fusion studied here, the latter genetic lesion is predominantly a prenatal and presumed initiating event[7]. It is coupled with a modest number (3–6) of recurrent, genomic copy number alterations (CNA)[8]. These accrue as secondary and, most likely, postnatal lesions[9] in genes that, predominantly, regulate the cell cycle or B-cell differentiation[8].

Subclonal diversity of genotypes in ALL

We initially selected 60 cases of *ETV6–RUNX1*-positive ALL and in which *ETV6* was also deleted (15–85% of cells) as detected by fluorescence *in situ* hybridization (FISH). Of these, 30 were further selected (see Supplementary Table 1) that also had (by FISH) deletion of *PAX5* ($n = 15$) or *CDKN2A* (also called *p16*) ($n = 12$) or deletions of both *PAX5* and *CDKN2A* ($n = 3$) in at least 10% of cells. All 30 cases were then scrutinized using a multiplexed combination of distinctive fluorochrome-labelled bacterial artificial chromosome (BAC) probes. Two-hundred cells with the *ETV6–RUNX1* fusion signal (that is, the reference founder mutation present in all leukaemic cells) were evaluated for each case and each individual cell designed an allele status (that is, mono- or bi-allelic deletion) for *ETV6* and *PAX5* (three colour) or *ETV6* and *CDKN2A* (three colour) or *ETV6*, *PAX5* and *CDKN2A* (four colour). The use of an *ETV6–RUNX1* probe also allowed us to detect duplication of the fusion gene (in 15 out of 30 cases) or an extra copy of chromosome 21q (via *RUNX1* signal copy number; in 21 out of 30 cases). The latter is a common genetic abnormality in ALL and an assumed driver event[10]. Cutoff levels (%) for scoring genetically distinctive subclones were determined using normal blood controls and varied depending upon probe set combination. A threshold was set at 2% for cells with a single CNA (in addition to *ETV6–RUNX1* fusion) and 1% for cells with two or more CNA (see Methods and Supplementary Table 2).

Enumeration of CNA in individual cells in reference to *ETV6–RUNX1* fusion—the universal marker of all the leukaemic cells—allowed us to identify distinctive genetic signatures of subclones and their relative frequencies. From this we could infer the most likely evolutionary or ancestral relationships between the subclones and derive a clonal architecture.

The genetic architectures that were observed were very diverse. The simplest of these genetic architectures that we identified (in 6 cases) involved two or three subclones that could be aligned in a linear sequence (Fig. 1a); however, these were cases with the lowest complement of CNA (only two or three out of the possible total of the seven pre-selected) in addition to the *ETV6–RUNX1* fusion. All

[1]Section of Haemato-Oncology, The Institute of Cancer Research, Sutton SM2 5NG, UK. [2]MRC Molecular Haematology Unit, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DS, UK. [3]Paediatric Malignancy Unit, Great Ormond Street Hospital & UCL Institute of Child Health, London WC1N 3JH, UK. [4]Leukaemia Research Cytogenetics Group, Northern Institute for Cancer Research, Newcastle University, Newcastle upon Tyne NE1 4LP, UK. †Present address: University College London Cancer Institute, London WC1E 6BT, UK.

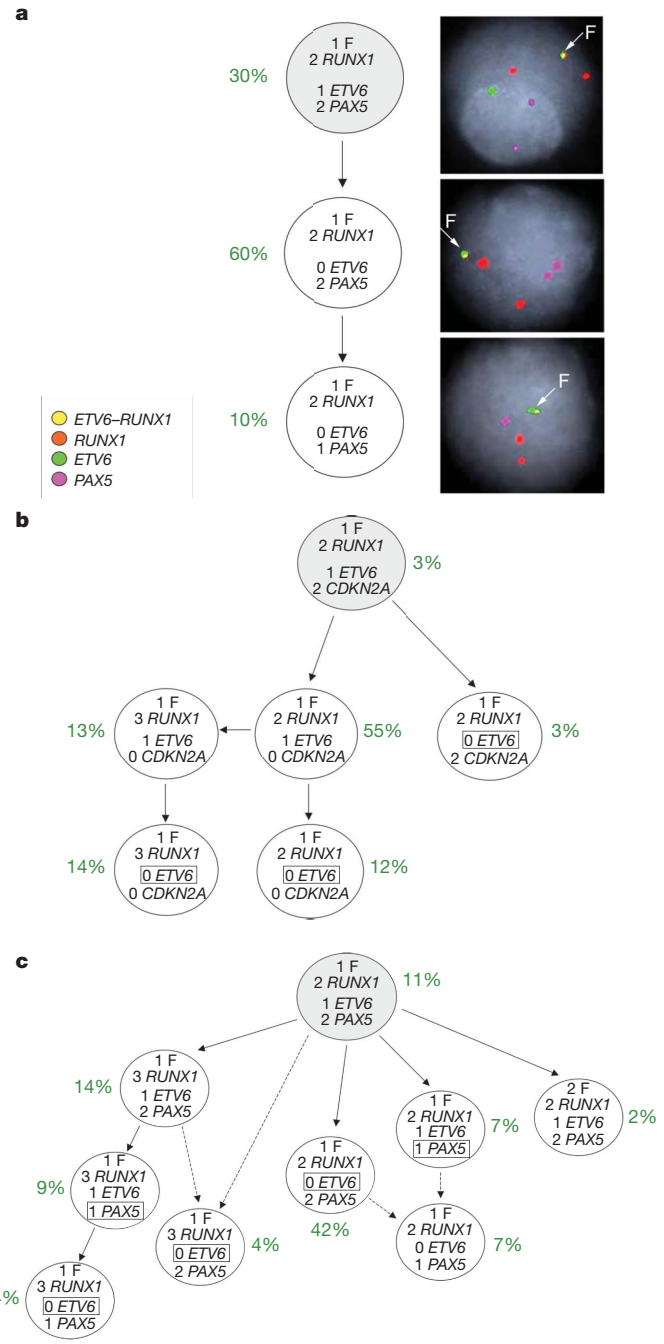


Figure 1 | Examples of subclonal architecture in ALL. **a**, Apparent linear architecture, with three clones (patient no. 13). Representative FISH images on the right show examples of each subclone. **b**, Moderately complex architecture with five subclones (patient no. 8). Loss of the untranslocated *ETV6* allele occurs independently in three separate subclones (boxes). **c**, Complex architecture with eight subclones (patient no. 16). *PAX5* deletions occur independently in two separate subclones (boxes). Arrows indicate probable (or most likely) ancestral derivation of subclones; dashed arrows indicate possible (or alternative) origins of subclones. F, yellow signal, *ETV6–RUNX1* fusion gene; 2 *RUNX1*, two red signals (one large, one small) corresponding to one normal *RUNX1* allele, and one small remnant generated from disruption of *RUNX1* allele involved in the gene fusion; *ETV6*, green signal corresponding to the normal (untranslocated) *ETV6* allele; *PAX5* (or *CDKN2A*), pink signal.

other 24 cases had a more marked subclonal heterogeneity with up to ten subclones related via a branching ancestral tree (Fig. 1b, c). Figure 1a–c illustrates examples of the clonal architectures observed (all other cases are depicted in Supplementary Fig. 2).

Inspection of clonal genotypes reveals some previously unrecognized features. It is apparent that the common or highly recurrent CNA are not acquired in any preferential order, indicating that their potency as oncogenic mutations may not be contingent upon (or epistatic to) other CNA. Subclones with the highest number of CNA, positioned 'terminally' in the branching architecture, were not necessarily numerically dominant (for example, all three cases illustrated in Fig. 1). Unexpectedly, CNA involving the same gene could be simultaneously present in distinct subclones and must therefore arise more than once, independently. *ETV6* was independently deleted two to three times in 14 of the 30 cases (see Fig. 1b, c), *PAX5* deleted two or three times in 8 out of 18 cases (see Fig. 1c) and *CDKN2A* deleted two times in 4 of 15 cases. This raises interesting mechanistic questions and suggests that these lesions are not only selected on the basis of clonal advantage but may be targeted for DNA-level breakage. One possible mechanism is via off-target effects of RAGS or AID[11–13].

Immunophenotypes and genetic diversity

There is a spectrum of early B-lineage differentiation-linked immunophenotypic signatures in ALL[14] and evidence has been presented indicating that cells with several different antibody-defined phenotypes may have leukaemia propagating activity *in vivo*[15]. We analysed the genetic heterogeneity of cells flow sorted on the basis of their expression of CD34 (immature lineage marker) or CD20 (more mature B lineage marker). Sorted populations had similarly complex genetic architectures (Supplementary Fig. 3).

Cells with the immunophenotype $CD34^{+}CD38^{-/low}CD19^{+}$ appear, so far, to be unique to ALL[16]. We previously found this pro-B/stem population, possibly non-activated or quiescent ($CD38^{-}$), to be significantly enriched in ALL propagating cells when assayed in the NOD/SCID strain of mice[17]. When purified by cell sorting, these cells (from patient no. 7) had similar genetic complexity to the bulk leukaemic population (Supplementary Fig. 3).

Clonal architecture in ALL is dynamic

These descriptions of subclonal, genetic profiles in ALL are snapshots taken at a particular time point, that is, at diagnosis. It is likely that subclonal diversity and the relative dominance of subclones varies continuously with the development and progression of disease. ALL rarely has an identified prodromal phase but occasionally (~2%) patients with ALL have an aplastic, pre-leukaemic phase a few months before a diagnosis of leukaemia[18]. We previously described one such patient with *ETV6–RUNX1*$^{+}$ ALL[19]. The diagnostic ALL cells had *ETV6–RUNX1* fusion but no *ETV6* deletion. Single nucleotide polymorphism (SNP) array screening revealed multiple deletions including *BTG1* and 11q and gain of chromosome X. We compared the clonal genotypes of cells from this patient at these two time points, spread some 7 months apart, and observed a marked shift in clonal architecture (Fig. 2a). The subclones dominating the aplastic, pre-malignant phase were relegated to minor, intermediary subclones in the overt leukaemic phase, with dominance of progeny clones that had homozygously deleted *CDKN2A*. A second patient with a prodromal, aplastic phase some 3 months before a diagnosis of ALL also showed a shift in subclonal dominance (Supplementary Fig. 4).

Treatment and subsequent relapse in ALL reorders the spectrum of genetic abnormalities detected by single gene probing[20] or SNP arrays[21] reflecting the probable selection of distinct subclones as a basis of relapse[21]. For five of the thirty selected patients (numbers 6, 9, 28, 29 and 30), we had matched diagnosis and relapse cells available for multiplexed FISH analysis. Clonal architecture at relapse was different from that at diagnosis. The comparative genetic profiles of subclones allowed us to identify the most likely subclone giving rise to the relapse, although this attribution was not unambiguous (Fig. 2b and Supplementary Fig. 2). Relapse seems to derive from either major or minor clones at diagnosis as previously suggested[21] but with a suggestion that more than one subclone might contribute to relapse

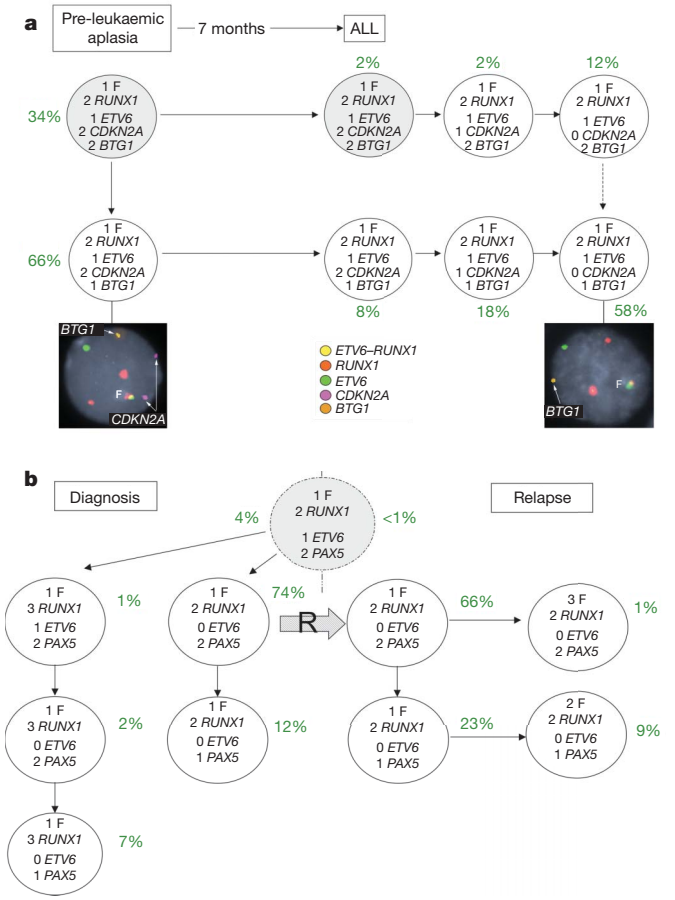


Figure 2 | Changes in clonal architecture in ALL. **a**, Scoring for *ETV6–RUNX1* fusion and simultaneous deletion of *CDKN2A* and *BTG1* in blast cells at diagnosis of ALL and in bone marrow taken 7 months earlier during a pre-leukaemic aplasia phase. During the pre-leukaemic aplasia phase no deletion of *CDKN2A* was present, although several other copy number abnormalities were present including 11q, 15q, 5q and *BTG1* gene deletions as well as gain of Xq (ref. 19). At diagnosis of ALL some 7 months later, the predominant clone contained homozygous deletion of *CDKN2A*. Only clones above the cutoffs are shown. Representative, four-colour FISH pictures for the dominant clones during the aplasia and leukaemic phases are shown at the bottom. **b**, An example of relapse originating from a major clone at diagnosis (patient no. 9) (other matched relapse cases for patients 6, 28, 29, 30 are in Supplementary Fig. 2). R, probable clonal origin of relapse.

(for example, patients 6 and 30; Supplementary Fig. 2). The data also indicate that the dominant subclone in relapse itself continues to genetically diversify, in some cases acquiring genetic lesions in the same gene (or chromosome region) as observed in primary, diagnostic subclones. This, along with previous observations on distinctive *ETV6* deletions in relapse versus diagnosis[20], provides further evidence for reiterative CNA. The patterns of genetic diversity observed in relapse indicate that genetically distinct leukaemic propagating cells can survive chemotherapy and provide a reservoir for relapse and further diversification.

Genetic diversity of propagating cells in ALL

Within the genetic architecture of ALL, it cannot be assumed that all identified subclones are self-sustaining and propagated by cells with extensive self-renewing capacity[5]. As in evolutionary speciation, it is likely that some branches or subclones are long-lived whereas others are dead ends or out-competed. Nevertheless, the architectural patterns that we observed suggested the possibility that propagating cells for ALL might also have variegated genetics and that this should be demonstrable via serial transplantation in immunodeficient mice. We transplanted, intra-tibially, varying numbers (2×10^3–10^6) of unfractionated or immunophenotypically flow-sorted leukaemic cells into pre-irradiated NOD/SCID IL2Rγ^{null} mice. Expanded leukaemic populations were re-transplanted into secondary recipient mice as a validation of self-renewal capacity. We compared the genetic signatures of the cell populations that emerged by successful regeneration *in vivo*, from first and secondary transplants, with those in the original diagnostic sample. Mice with regenerated ALL had significant proportions (3.1 to 93.5 av. 59.4; Supplementary Tables 3 and 4) of human haematopoietic (CD45$^+$) cells in the marrow and large, pale spleens (Supplementary Fig. 5b, c). Effectively, all (>99%) human CD45$^+$ cells were leukaemic with the *ETV6–RUNX1* fusion (Supplementary Fig. 5d). Genetic analysis was carried out on cells harvested from bone marrow but when assessed, spleen provided the same result (Supplementary Table 4).

Leukaemic regeneration *in vivo* was observed consistently in both unfractionated populations and in fractions defined immunophenotypically as CD34$^+$CD38$^{-/low}$CD19$^+$ and CD34$^+$CD38$^+$CD19$^+$ (Supplementary Tables 3 and 4). This accords with previous evidence[15] that propagating cells in B-cell precursor ALL are not restricted to one immunophenotypic compartment.

In all 24 mice with primary or secondary leukaemic regeneration, several genetically distinct subclones were present, the patterns of which reflected the diversity of subclones identified in the original diagnostic sample (Figs 3 and 4 and Supplementary Tables 3 and 4). The leukaemic cells regenerated in secondary transplants were compared to pre-transplant primary cells by high-resolution SNP arrays. For patient no. 3, these data confirmed subclonal loss of *CDKN2A*, subclonal gain of chromosome 21 and loss of one copy of *ETV6* in

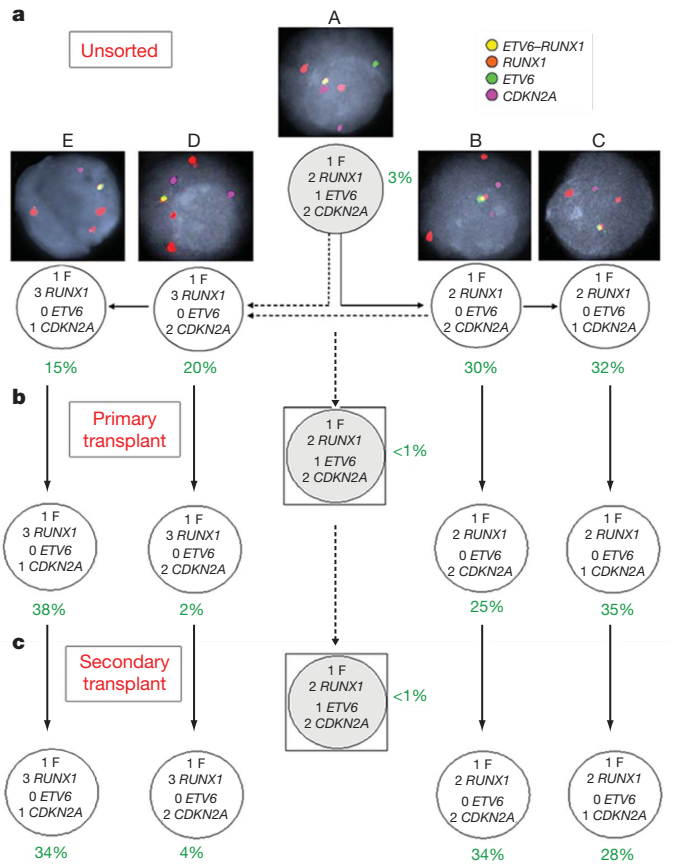


Figure 3 | Genetics of cells propagating NOD/SCID IL2Rγ^{null} mice. Leukaemic cells from patient no. 3 before injection (**a**), after primary transplantation (**b**; mouse 1, Supplementary Table 3) and after secondary transplantation (**c**; mouse 2, Supplementary Table 3). Representative FISH images are shown of the four subclones (B–E) and the putative pre-leukaemic cell (A) at the top in **a**. Boxed cell, below significance threshold.

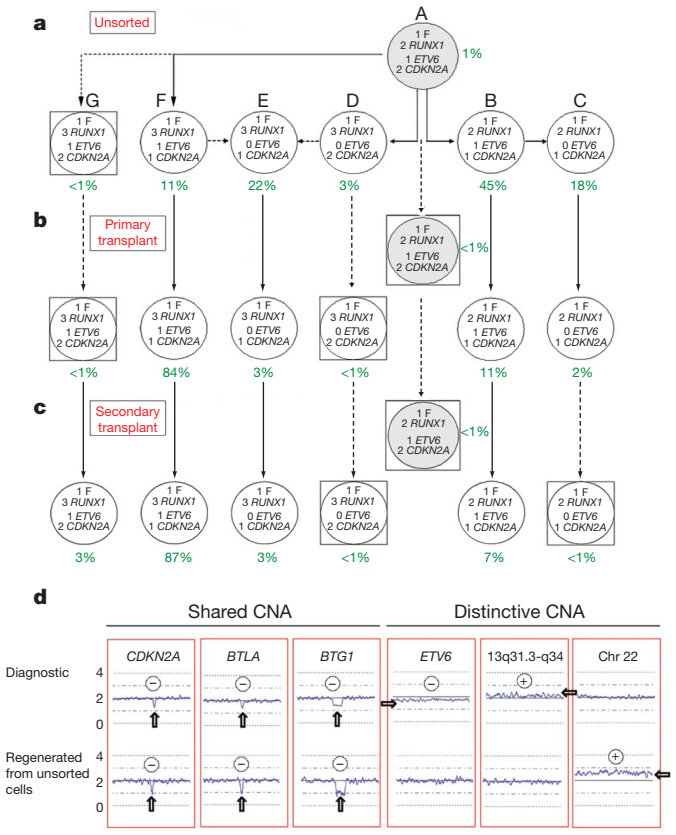


Figure 4 | Shift in clonal architecture of ALL after *in vivo* NOD/SCID IL2Rγ^{null} transplantation. Clonal architecture of unsorted cells from patient no. 7 before injection (**a**), after primary transplantation (**b**; mouse 1, Supplementary Table 4) and after secondary transplantation (**c**; mouse 2, Supplementary Table 4). **d**, Summary of SNP data on diagnostic versus secondary transplant ALL cells. The blue lines indicate the mean copy number plot of five contiguous SNPs. Middle line, normal diploid copy number. Two fractions of DNA have been examined using 500K SNP arrays: diagnostic DNA (unsorted cells from patient no. 7 before injection) and DNA from leukaemic cells regenerated in mice (mouse 2, Supplementary Table 4). Deletions of *CDKN2A*, *BTLA* and *BTG1* are present in both diagnostic and regenerated samples, whereas deletion of *ETV6*, subclonal gain of the 13q31.3-q34 region and gain of chromosome 22 are distinctive between the two samples. Plus and minus symbols indicate gains or losses of genetic regions, respectively. Arrows highlight CNA.

both diagnostic and regenerated samples (Supplementary Fig. 6 and Supplementary Table 3).

The genetic profiles of regenerated leukaemias revealed variable potency of genetically distinct subclones. For patient no. 3 (Fig. 3 and Supplementary Table 3), four subclones read-out in primary and secondary transplants with three being co-dominant. The putative pre-leukaemic clone (with *ETV6–RUNX1* fusion only) did not regenerate. With unfractionated cells from patient no. 7 (Fig. 4 and Supplementary Table 4), four of six subclones regenerated upon secondary transplantation. One of these (subclone F in Fig. 4) was dominant despite being a minor (11%) subclone in the initial diagnostic sample. SNP arrays were used to compare the primary diagnostic sample (patient no. 7) versus the regenerated (secondary) transplant leukaemias. These revealed that both leukaemic populations had *BTLA* and *BTG1* deletions (Fig. 4d) in addition to the CNA in *CDKN2A*, *ETV6* and chromosome 21. The regenerated leukaemia had an additional chromosome 22 that appeared to be absent from the initial, diagnostic cell population (Fig. 4d). This was further investigated by FISH using a chromosome-22-specific BAC probe. This confirmed the SNP array data with the majority of cells (with three *RUNX1* and one *ETV6* signals) in the regenerated leukaemia having an extra chromosome 22 signal (Supplementary Fig. 7). No cells with an extra chromosome 22 signal were detectable in the initial diagnostic sample in accord with the SNP array data. However, because clone F was dominant in all 9 of 9 mice transplanted with the same cell population (Supplementary Table 4), we assume that a minor subclone of clone F with the extra chromosome 22 was present in the diagnostic sample but at a $<1\%$ frequency. Variable competitive potency of subclonal regeneration was also seen in mice injected with immunologically fractionated cells (Supplementary Table 4 and Supplementary Figs 7 and 8).

These data are indicative of additional genetic complexity of subclones and their propagating cells. Moreover, they indicate that distinctive genotypes are associated, functionally, with variable competitive regeneration *in vivo*.

Discussion

Cancer development at the cellular level is widely regarded as a Darwinian evolutionary process involving 'natural selection' of genetically variant cells in the context of a complex micro-environmental ecology[22–24]. Mutational and phenotypic diversity between cells is, in principle, fundamental to this process. Moreover, driver mutations can be expected to have maximal selective currency when present in cells with self-renewing functionality.

Evidence for intraclonal genetic diversity in cancer has been provided by chromosome karyotype[25], by genetic analysis of multi-focal (but monoclonal) cancers[26], by FISH-based screening of tissue sections[27–29] or immuno-selected cells[30], by the molecular probing of multiple small biopsies[31] or of micro-dissected tissue[32–34] and, recently, by sector-ploidy profiling[35]. Small numbers of individual circulating tumour cells have also been scrutinized for their divergent genetic profiles[36,37]. These studies collectively testify that contemporaneous intraclonal genetic heterogeneity is commonplace and, in some cases at least, the degree of clonal diversity is predictive of disease progression[31]. Most of these data derive from epithelial carcinomas with complex genetic profiles, coupled, in most cases, to genetic instability. In such cases the historical timing and sequence of critical or driver mutational events is effectively buried and clonal architecture could be extremely complex unless clonal dominance occurs.

A common assumption for both leukaemias and cancer in general, based on the original evolutionary model of ref. 22, is that progression of disease and predominant genetic profiles reflect sequentially dominant clones and an essentially linear dynamic. Our data (summarized in Supplementary Fig. 1) suggest dynamic patterns of subclonal development and ancestral relationships that are nonlinear with a variable branching architecture. Patterns of genetic diversity in other cancers—assessed by single cell or ploidy sorted cell comparative genomic hybridization (CGH)[35,38,39], oncogenically neutral microsatellite markers[31,40] or deep-sequenced IGH gene rearrangements[41]—also indicate nonlinear, branching clonal trajectories. Collectively, these data indicate that cancer has a cellular and genetic architecture reminiscent of Darwin's iconic evolutionary tree (or bush) diagram depicting speciation[42].

The extent of genetic variegation in subclones that we detect must be a significant underestimate. We screened for a limited number of pre-selected CNA, which means that other CNA plus any sequence-based driver mutations present will not have been registered. Moreover, antecedent or intermediary subclones, below the 1–2% frequency which we set as a threshold, were identified with more extensive screening in several cases (see Supplementary Fig. 2 patients 2, 5, 15, 18, 20, 6). Identifying the full complexity of subclonal architecture and genetic diversity in ALL (and other cancers) will ultimately require whole-genome analysis at the single cell level.

Our data provide the first direct evidence for genetic diversity of cancer propagating cells within individual patients. The consistent patterns of subclonal regeneration in mice (that is, in different mice injected with the same cellular inoculum) suggest variable capacity intrinsically associated with the distinct genotypes of propagating

cells. The competitive potency of particular subclones observed, however, may to some extent reflect selective pressures exerted by regenerative stress in a murine tissue environment. Natural clonal selection in patients might produce different outcomes.

It will be important to assess if genetic diversity of propagating cells holds true for other types of leukaemia and cancer in general. If it does, then there would be significant implications for both the cancer-stem-cell concept itself and for the therapeutic targeting of such cells. The original model of a distinct, hierarchically positioned subpopulation of cancer stem cells[5] has proved contentious in both ALL[15] and other cancers[43–48]. It has been suggested that the NOD/SCID *in vivo* readout for human cancer stem cell may, at least for some cancers, simply register dominant subclones[43,48]. Or, alternatively, that cancer stem cells exist but evolve over time[44,45]. We have previously documented that 'pre-leukaemic' and overt leukaemia propagating cells in *ETV6–RUNX1*-positive ALL, although clonally related by descent, are distinctive in IgH rearrangements and phenotype[17]. Our current data fit best with what we refer to as a 'back to Darwin' model for cancer propagating cells and resultant clonal architecture[46]. In this, cells with self-renewing properties have variegated genotypes providing the units of selection in the evolutionary diversification and progression of disease. Both sequential and concurrent genotypic variation in propagating cells occur in ALL and, we predict, are likely to do so in other cancers, providing a rich substrate for progression of disease. Although it has yet to be evaluated, it is likely that genetic diversity of cancer propagating cells will be associated with both frequency variation and diversity of functional properties, for example, differentiation status, niche occupancy, quiescence and drug or irradiation sensitivity. This may help to explain some of the inconsistencies and controversies in the cancer-stem-cell field[44, 47,48,49]. Genetic diversity in cancer varies in extent with stage of disease[28,33], probably reflecting the impact of intraclonal competition and ecological bottlenecks. Single cells might negotiate very stringent bottlenecks but the genetic profiles that we observed in relapsed ALL and as recorded in, for example, prostate cancer metastases[37,50] indicate continued diversification of propagating cells and dominant or therapy-resistant subclones.

This perspective contrasts with the unidimensional or flat (albeit very complex) genetic landscapes of cancer implied in portraits derived from whole-genome scans. This architectural distinction may be of some clinical consequence. Targeted therapy, if directed at mutant molecules, may have limited efficacy if the targets themselves are not initiating lesions but secondary mutations segregated in subclones, even when the latter appear dominant. Genetic variegation of cancer propagating cells may represent a significant roadblock to effective therapy.

METHODS SUMMARY

Archival methanol:acetic-acid-fixed cytogenetic pellets from patients with *ETV6–RUNX1* fusion-gene-positive ALL were obtained from several UK hospitals, with local ethical review committee approval (CCR 2285, Royal Marsden Hospital NHS Foundation Trust). The clinical and cytogenetic data on these patients are given in Supplementary Table 1. Interphase FISH was performed as previously described[9,19].

In each case, at least 200 nuclei were scored for the presence of the *ETV6–RUNX1* fusion gene in combination with hemizygous or homozygous deletion of *ETV6*, *RUNX1*, *PAX5*, *CDKN2A* and *BTG1* and 11q, as well as gain of *RUNX1* and duplication of the *ETV6–RUNX1* fusion gene. Controls included the scoring of residual normal cells within the diagnostic sample and scoring leukaemic cells with probes hybridizing to irrelevant oncogenes (*BCR*, *ABL*) (see main Methods and Supplementary Figs 9 and 10).

NOD/SCID IL2Rγ^{null} mice that lack any B, T and natural killer cell activity were bred and maintained under sterile conditions in accordance with Home Office regulations. Transplantation of cells was by intra-tibial injections in 7–14-week-old mice after 250 cGy irradiation. Peripheral engraftment was assessed at 9–10 weeks after transplantation and if >2% mice were killed. Further analysis included the assessment of bone marrow/spleen engraftment, FISH analysis, histological analysis and serial transplantation. For serial transplantations, recovered bone marrow cells were stained with human CD45 to detect human engraftment. An equivalent of 2×10^3 to 2×10^5 human cells was transplanted by intra-tibial injections.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 20 May; accepted 27 October 2010.
Published online 15 December 2010.

1. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
2. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).
3. Fox, E. J., Salk, J. J. & Loeb, L. A. Cancer genome sequencing—an interim analysis. *Cancer Res.* **69,** 4948–4950 (2009).
4. Marusyk, A. & Polyak, K. Tumor heterogeneity: causes and consequences. *Biochim. Biophys. Acta* **1805,** 105–117 (2010).
5. Dick, J. E. Stem cell concepts renew cancer research. *Blood* **112,** 4793–4807 (2008).
6. Klein, C. A. *et al.* Genetic heterogeneity of single disseminated tumour cells in minimal residual cancer. *Lancet* **360,** 683–689 (2002).
7. Greaves, M. F. & Wiemels, J. Origins of chromosome translocations in childhood leukaemia. *Nature Rev. Cancer* **3,** 639–649 (2003).
8. Mullighan, C. G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446,** 758–764 (2007).
9. Bateman, C. M. *et al.* Acquisition of genome-wide copy number alterations in monozygotic twins with acute lymphoblastic leukemia. *Blood* **115,** 3553–3558 (2010).
10. Loncarevic, I. F. *et al.* Trisomy 21 is a recurrent secondary aberration in childhood acute lymphoblastic leukemia with *TEL/AML1* fusion. *Genes Chromosom. Cancer* **24,** 272–277 (1999).
11. Kitagawa, Y. *et al.* Prevalent involvement of illegitimate V(D)J recombination in chromosome 9p21 deletions in lymphoid leukemia. *J. Biol. Chem.* **277,** 46289–46297 (2002).
12. Mullighan, C. G. *et al. BCR-ABL1* lymphoblastic leukaemia is characterized by the deletion of Ikaros. *Nature* **453,** 110–114 (2008).
13. Feldhahn, N. *et al.* Activation-induced cytidine deaminase acts as a mutator in *BCR-ABL1*-transformed acute lymphoblastic leukemia cells. *J. Exp. Med.* **204,** 1157–1166 (2007).
14. van Dongen, J. J. M., Szczepanski, T. & Adriaansen, H. J. in *Leukemia* (eds Henderson, E. S., Lister, T. A. & Greaves, M. F.) 85–129 (Saunders, 2002).
15. le Viseur, C. *et al.* In childhood acute lymphoblastic leukemia, blasts at different stages of immunophenotypic maturation have stem cell properties. *Cancer Cell* **14,** 47–58 (2008).
16. Castor, A. *et al.* Distinct patterns of hematopoietic stem cell involvement in acute lymphoblastic leukemia. *Nature Med.* **11,** 630–637 (2005).
17. Hong, D. *et al.* Initiating and cancer-propagating cells in *TEL-AML1*-associated childhood leukemia. *Science* **319,** 336–339 (2008).
18. Breatnach, F., Chessells, J. M. & Greaves, M. F. The aplastic presentation of childhood leukaemia: a feature of common-ALL. *Br. J. Haematol.* **49,** 387–393 (1981).
19. Horsley, S. W. *et al.* Genetic lesions in a preleukemic aplasia phase in a child with acute lymphoblastic leukemia. *Genes Chromosom. Cancer* **47,** 333–340 (2008).
20. Zuna, J. *et al. TEL* deletion analysis supports a novel view of relapse in childhood acute lymphoblastic leukemia. *Clin. Cancer Res.* **10,** 5355–5360 (2004).
21. Mullighan, C. G. *et al.* Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science* **322,** 1377–1380 (2008).
22. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194,** 23–28 (1976).
23. Gatenby, R. A. & Vincent, T. L. An evolutionary model of carcinogenesis. *Cancer Res.* **63,** 6212–6220 (2003).
24. Merlo, L. M. F., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nature Rev. Cancer* **6,** 924–935 (2006).
25. Teixeira, M. R. *et al.* Karyotypic comparisons of multiple tumorous and macroscopically normal surrounding tissue samples from patients with breast cancer. *Cancer Res.* **56,** 855–859 (1996).
26. Takahashi, T. *et al.* Clonal and chronological genetic analysis of multifocal cancers of the bladder and upper urinary tract. *Cancer Res.* **58,** 5835–5841 (1998).
27. Cottu, P. H. *et al.* Intratumoral heterogeneity of HER2/neu expression and its consequences for the management of advanced breast cancer. *Ann. Oncol.* **19,** 596–597 (2008).
28. Park, S. Y. *et al.* Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *J. Clin. Invest.* **120,** 636–644 (2010).
29. Clark, J. *et al.* Complex patterns of *ETS* gene alteration arise during cancer development in the human prostate. *Oncogene* **27,** 1993–2003 (2008).
30. Shipitsin, M. *et al.* Molecular definition of breast tumor heterogeneity. *Cancer Cell* **11,** 259–273 (2007).
31. Maley, C. C. *et al.* Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature Genet.* **38,** 468–473 (2006).
32. Aubele, M. *et al.* Intratumoral heterogeneity in breast carcinoma revealed by laser-microdissection and comparative genomic hybridization. *Cancer Genet. Cytogenet.* **110,** 94–102 (1999).
33. Boland, C. R. *et al.* Microallelotyping defines the sequence and tempo of allelic losses at tumour suppressor gene loci during colorectal cancer progression. *Nature Med.* **1,** 902–909 (1995).

34. Geyer, F. C. *et al.* Molecular analysis reveals a genetic basis for the phenotypic diversity of metaplastic breast carcinomas. *J. Pathol.* **220,** 562–573 (2010).
35. Navin, N. *et al.* Inferring tumor progression from genomic heterogeneity. *Genome Res.* **20,** 68–80 (2010).
36. Stoecklein, N. H. *et al.* Direct genetic analysis of single disseminated cancer cells for prediction of outcome and therapy selection in esophageal cancer. *Cancer Cell* **13,** 441–453 (2008).
37. Attard, G. *et al.* Characterization of *ERG, AR* and *PTEN* gene status in circulating tumor cells from patients with castration-resistant prostate cancer. *Cancer Res.* **69,** 2912–2918 (2009).
38. Klein, C. A. & Stoecklein, N. H. Lessons from an aggressive cancer: evolutionary dynamics in esophageal carcinoma. *Cancer Res.* **69,** 5285–5288 (2009).
39. Kuukasjärvi, T. *et al.* Genetic heterogeneity and clonal evolution underlying development of asynchronous metastasis in human breast cancer. *Cancer Res.* **57,** 1597–1604 (1997).
40. Tsao, J.-L. *et al.* Colorectal adenoma and cancer divergence. Evidence of multilineage progression. *Am. J. Pathol.* **154,** 1815–1824 (1999).
41. Campbell, P. J. *et al.* Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl Acad. Sci. USA* **105,** 13081–13086 (2008).
42. Barrett, P. H. *et al.* (eds.) *Charles Darwin's Notebooks*, 1836–1844 (Cambridge Univ. Press, 1987).
43. Adams, J. M. & Strasser, A. Is tumor growth sustained by rare cancer stem cells or dominant clones? *Cancer Res.* **68,** 4018–4021 (2008).
44. Visvader, J. E. & Lindeman, G. J. Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. *Nature Rev. Cancer* **8,** 755–768 (2008).
45. Rosen, J. M. & Jordan, C. T. The increasing complexity of the cancer stem cell paradigm. *Science* **324,** 1670–1673 (2009).
46. Greaves, M. Cancer stem cells: back to Darwin? *Semin. Cancer Biol.* **20,** 65–70 (2010).
47. Maenhaut, C., Dumont, J. E., Roger, P. P. & van Staveren, W. C. G. Cancer stem cells: a reality, a myth, a fuzzy concept or a misnomer? An analysis. *Carcinogenesis* **31,** 149–158 (2010).
48. Shackleton, M., Quintana, E., Fearon, E. R. & Morrison, S. J. Heterogeneity in cancer: cancer stem cells versus clonal evolution. *Cell* **138,** 822–829 (2009).
49. Polyak, K. Breast cancer: origins and evolution. *J. Clin. Invest.* **117,** 3155–3163 (2007).
50. Liu, W. *et al.* Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nature Med.* **15,** 559–565 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work is supported by specialist programme grants from The Kay Kendall Leukaemia Fund (M.G.) and Leukaemia & Lymphoma Research (M.G., T.E.) and a Deutsche Forschungsgemeinschaft fellowship LU 1474/1-1 (to C.L.). T.E. and C.L. acknowledge support from the Oxford BRC.

Author Contributions K.A. carried out the FISH analyses. C.L. and Y.G. conducted the *in vivo* experiments. C.M.B. analysed the SNP array and FISH analysis of the patient with aplasia and ALL. S.M.C. and I.T. performed cell immunostaining and sorting. F.W.v.D. provided SNP array data. H.K., A.V.M. and J.S. provided patient data and samples. T.E. advised on design and interpretation of *in vivo* experiments. L.K. supervised the FISH studies. L.K. and M.G. designed the overall study. M.G. wrote the paper with critical input from T.E., L.K. and other authors.

Author Information SNP array data have been deposited in the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession code GSE24412. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.G. (mel.greaves@icr.ac.uk).

METHODS

Interphase fluorescence *in situ* hybridization (FISH). Archival methanol:acetic-acid-fixed cytogenetic pellets from patients with *ETV6–RUNX1* fusion-gene-positive ALL were obtained from several UK hospitals, with local ethical review committee approval (CCR 2285, Royal Marsden Hospital NHS Foundation Trust). Interphase FISH for the *ETV6–RUNX1* fusion gene was performed using a commercial LSI *TEL-AML1* extra signal (ES) probe (Vysis, Abbott Laboratories Ltd) according to the manufacturers' instructions. This probe set contains a 350-kb probe for the 5′ end of *ETV6* (exons 1–4) and a 500-kb probe covering the entire *RUNX1* gene. The FISH signal pattern for the *ETV6–RUNX1* fusion-gene-positive cells using the Vysis probe is two red (one large, one small *RUNX1* signals), one green (*ETV6* allele not involved in the translocation), one red/green (yellow) fusion signal corresponding to the *ETV6–RUNX1* fusion gene. Bacterial artificial chromosome (BAC) or fosmid probes for the *PAX5, CDKN2A*, *BTG1*, *TBL1XR1* genes, 11q and other regions of interest were obtained from the BACPAC Resource Centre, Children's Hospital, Oakland Research Institute (http://bacpac.chori.org). These were labelled by nick translation with biotin-16-dUTP or digoxigenin-11-dUTP (Roche) and hybridized in combination with the *ETV6–RUNX1* ES probe. FISH was performed by standard protocols[9,19] and labelled probes detected with streptavidin-Cy5 (bioytinylated probes) and (1) monoclonal anti-digoxigenin (Sigma), (2) horse anti-mouse IgG-Texas red (Vector Laboratories) and (3) goat anti-horse IgG-Texas Red (Jackson Immunochemicals) (for digoxigenin-labelled probes). Fluorescent signals were viewed using an Olympus AX2 fluorescence microscope equipped with narrow bandpass filters for DAPI, FITC, Spectrum orange, Texas red and Cy5. Images were captured and analysed using a charge-coupled device (Photometrics) and SmartCapture 3 software version 3.0.4 (Digital Scientific).

Establishing cutoff levels. In each case, at least 200 nuclei were scored for the presence of the *ETV6–RUNX1* fusion gene in combination with hemizygous or homozygous deletion of *ETV6*, *RUNX1*, *PAX5*, *CDKN2A* and *BTG1* and 11q, as well as gain of *RUNX1* and the *ETV6–RUNX1* fusion gene. Diagnostic slides from 26 cases were assessed for hybridization efficiency by scoring the residual normal (*ETV6–RUNX1* fusion negative) cells on the same slide (see Supplementary Fig. 9). The percentage of these cells with loss of a single *CDKN2A* or *PAX5* signal was 0–3% (mean = 1%). As a further control, we hybridized a subset ($n = 11$) of *ETV6–RUNX1* fusion-gene-positive cases with uninvolved oncogene probes *BCR* and *ABL* (see Supplementary Fig. 10). The percentage of cells with the expected normal signal pattern (two red, two green) was 96–99% (mean = 98%). Cutoff levels for each probe were established by three-colour FISH (test probe in combination with the *ETV6–RUNX1* ES probe) using three normal control peripheral blood slides per probe. As *ETV6* and *RUNX1* were scored on each slide, the values for these two probes were based on 12 slides in total. We used a cutoff = mean + 2× standard deviation. Cutoff levels for each probe (in combination with the *ETV6–RUNX1* fusion) were established using 3–12 normal control peripheral blood slides. The cutoff levels for three-colour FISH experiments are given in Supplementary Table 2. Cutoff levels for four-colour FISH were the same as above, except for Texas red probes. The cutoff for loss of one signal using a fourth probe detected with Texas red was 6.9%, because of spectral overlap between Texas red (used to detect digoxigenin-labelled probes) and Spectrum orange (used to label *RUNX1* in the commercial *ETV6–RUNX1* ES probe). However, in most cases we used three-colour FISH as a cross-check to infer whether clones below this cutoff were real (Supplementary Fig. 2). Only fusion-gene-positive cells that also showed the small extra red signal (generated by disruption of *RUNX1*) were used to calculate the relative frequencies of the various subclones.

Genome mapping analysis. Mapping analysis was performed using 500 ng of tumour DNA. DNA was prepared according to manufacturer's instructions using the GeneChip mapping 500K assay protocol for hybridization to GeneChip Mapping 250K Nsp and Sty arrays (Affymetrix). Briefly, genomic DNA was digested in parallel with restriction endonucleases NspI and StyI, ligated to an adaptor, and subjected to polymerase chain reaction (PCR) amplification with adaptor-specific primers. The PCR products were digested with DNaseI and labelled with a biotinylated nucleotide analogue. The labelled DNA fragments were hybridized to the microarray, stained by streptavidin-phycoerythrin conjugates, and washed using the Affymetrix Fluidics Station 450 then scanned with a GeneChip scanner 3000 7G.

Copy number and LOH analysis. SNP genotypes were obtained using Affymetrix GCOS software (version 1.4) to obtain raw feature intensity and Affymetrix GTYPE software (version 4.0) using the BRLMM algorithm to derive SNP genotypes. The samples were analysed using CNAG 3.0 (http://plaza.umin.ac.jp/genome), comparing tumour sample with unpaired control DNA to determine copy number and loss of heterozygosity (LOH) caused by imbalance[51]. The position of regions of LOH were identified using the University of California Santa Cruz (UCSC) Genome Browser, May 2004 Assembly (http://genome.ucsc.edu/cgi-bin/hgGateway).

Combined fluorescence immunophenotype and FISH. Bone marrow and spleen cells from leukaemic mice were cytospun and used for combined or triple-colour immunophenotype/FISH analysis as previously described[52]. Briefly, cells were air dried and fixed in acetone before incubating in primary biotinylated mouse anti-human CD45 (Clone F10-89-4) and detecting with Avidin-AMCA (Vector Laboratories). After antibody staining, the slides were hybridized with the Vysis *ETV6–RUNX1* ES fusion gene FISH probe as described above. Cells were viewed using a Zeiss Axioskop fluorescence microscope fitted with a dual bandpass FITC and rhodamine filter, as well as individual DAPI (for AMCA immunophenotype), FITC, rhodamine and Cy5. Images were captured using a charge-coupled device (Photometrics) and fluorescence signals merged and analysed using SmartCapture X software version 2.6.2 (Digital Scientific).

Cell separation, phenotyping and sorting. Total mononuclear cells were isolated by Ficoll gradient centrifugation and directly cryopreserved in DMSO for later use. After thawing dead cells were evaluated and excluded by FACS after staining with Hoechst 33258 (Invitrogen). For sorting CD34 and CD20 positive and negative subsets, samples were stained with either mouse anti-human CD34 (IgG_1, Dako) or CD20 (IgG_1, Southern Biotech) followed by anti-mouse IgG labelled with Pacific blue (Invitrogen). Nonspecific binding of antibody was assessed using mouse IgG_1 isotype control stained samples (Dako). Sorting was performed on a BD FACSAria with analysis by Becton Dickinson FACSDiVa software. Samples were gated on forward- and side-scatter plots for mononuclear cells and further gated in forward-scatter height versus area to exclude clumped cells. Before xeno-transplantation, some cells were stained with anti-CD19 PE (BD Pharmingen), CD34 FITC (BD Pharmingen) and CD38 APC (BD-Pharmingen). $CD34^+38^{-/low}CD19^+$ and pro-B $CD34^+CD38^+CD19^+$ cells were purified by flow cytometry (in this case, using MoFlo, Dako). Data acquisition and analysis were done with Summit (Dako) software. For multi-colour cell sorting 'fluorescence minus one' controls were used to determine positive and negative staining boundaries[53]. Human cells regenerating in mice were identified by staining with anti-CD45 PeCy7 (BD Pharmingen).

NOD/SCID mouse transplantation. NOD/SCID IL2Rγ^{null} mice that lack any B, T and natural killer cell activity were bred and maintained at the Weatherall Institute of Molecular Medicine animal facility in accordance with Home Office regulations. Animals were handled under sterile conditions. Transplantations of 2×10^3–10^6 cells were performed by intra-tibial injections in 7–14-week-old mice. Recipients received 250 cGy of total body irradiation before cell injection. Peripheral engraftment was assessed at 9–10 weeks after transplantation and if peripheral engraftment was >2% mice were killed. Further analysis included the assessment of bone marrow/spleen engraftment, FISH analysis, histological analysis and serial transplantation. For serial transplantations recovered bone marrow cells were stained with human CD45 to detect human engraftment. An equivalent of 2×10^3 to 2×10^5 human cells was transplanted by intra-tibial injections.

May–Grünwald Giemsa staining. The histological analysis of patient samples and mouse bone marrow was performed by May–Grünwald Giemsa staining of bone marrow smears, bone marrow cytospin preparations and spleen swabs. Slides were analysed on an Olympus BX60 microscope.

51. Nannya, Y. *et al.* A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res.* **65,** 6071–6079 (2005).
52. Kearney, L. & Colman, S. in *Methods in Molecular Biology. Leukemia. Methods and Protocols* Vol. 538 (ed. So, C. W. E.) 57–70 (Humana Press, 2009).
53. Maecker, H. T. & Trotter, J. Flow cytometry controls, instrument setup, and the determination of positivity. *Cytometry A* **69,** 1037–1042 (2006).

ARTICLE

doi:10.1038/nature09733

Evolution of human *BCR–ABL1* lymphoblastic leukaemia-initiating cells

Faiyaz Notta[1,2]*, Charles G. Mullighan[3]*, Jean C. Y. Wang[1,4], Armando Poeppl[1], Sergei Doulatov[1,2], Letha A. Phillips[3], Jing Ma[5], Mark D. Minden[4], James R. Downing[3] & John E. Dick[1,2]

Many tumours are composed of genetically diverse cells; however, little is known about how diversity evolves or the impact that diversity has on functional properties. Here, using xenografting and DNA copy number alteration (CNA) profiling of human *BCR–ABL1* lymphoblastic leukaemia, we demonstrate that genetic diversity occurs in functionally defined leukaemia-initiating cells and that many diagnostic patient samples contain multiple genetically distinct leukaemia-initiating cell subclones. Reconstructing the subclonal genetic ancestry of several samples by CNA profiling demonstrated a branching multi-clonal evolution model of leukaemogenesis, rather than linear succession. For some patient samples, the predominant diagnostic clone repopulated xenografts, whereas in others it was outcompeted by minor subclones. Reconstitution with the predominant diagnosis clone was associated with more aggressive growth properties in xenografts, deletion of *CDKN2A* and *CDKN2B*, and a trend towards poorer patient outcome. Our findings link clonal diversity with leukaemia-initiating-cell function and underscore the importance of developing therapies that eradicate all intratumoral subclones.

A widely accepted tenet of cancer biology is that most tumours arise from single cells and that multiple genetic alterations accumulate over time, resulting in transformation[1]. Historically, this process was considered to be a stepwise acquisition of new mutations, some of which provide a competitive growth advantage, resulting in successive rounds of clonal expansion with concomitant loss of earlier, less-fit clones[2]. In this model of tumour evolution, all clones are linearly related to each other. However, new genomic technologies are revealing a more complex clonal architecture in some cancers[3–6]. Analysis of chromosomal translocation breakpoints and CNA profiling in twins with *ETV6–RUNX1*-positive acute lymphoblastic leukaemia (ALL) showed that a pre-leukaemic clone is initiated *in utero* that expands, seeds both twins, and then evolves with different kinetics and CNA acquisition in each twin[7,8]. Genome-wide CNA profiling of paired diagnostic and relapse ALL samples has been particularly informative[9–11]. In most cases, the relapse clone shared only limited genetic identity with the predominant diagnostic clone and did not evolve from it. In the rest, the relapse clone was either identical or a direct evolutionary product of the diagnostic clone. These studies predicted the existence of an ancestral, pre-diagnostic clone that gave rise to at least two clonal lineages that evolved independently in many patients with ALL, with each clone acquiring different genetic aberrations: one clone giving rise to the dominant diagnostic clone and the other emerging as the predominant clone at relapse with the acquisition of additional CNA. These results indicate that tumour evolution may occur through a more complex branching model that gives rise to genetically distinct subclones at diagnosis that vary in aggressiveness and response to therapy[12]. However, proof of this model requires studies directly examining the functional properties of the cells in which genetic changes are found. Indeed, within leukaemias, non-proliferating or pre-apoptotic cells are often present and thus incapable of contributing to long-term clonal maintenance or to relapse. Therefore, functional studies of the cells responsible for driving leukaemic growth in patients must be combined with genetic approaches to identify genetically diverse subclones and their evolutionary ancestral precursors, and to show whether they possess biologically distinct growth properties.

Arguably the most important biological function a cancer cell can have is the ability to sustain clonal growth[13], a property best measured by tumour initiation assays in primary and secondary recipients. Indeed, some highly aggressive or metastatic tumours of mice and humans seem to be functionally homogeneous because almost every cell has tumour-initiating-cell capacity. However, most tumours appear to be functionally heterogeneous, as the tumour-initiating cells or leukaemia-initiating cells typically represent a minor fraction, although their frequency varies widely in syngeneic[14] or xenograft[15] recipients. It has been widely considered that intratumoral functional heterogeneity results from stochastic processes that influence cell growth but also from the variable behaviour of genetic subclones that arise through clonal evolution. The cancer stem cell model proposes an alternative explanation based on the hierarchical organization of the tumour clone where cancer stem cells are solely responsible for driving clonal growth and for therapeutic resistance. In the cancer stem cell model, tumour-initiating cells and cancer stem cells are synonymous and have the properties of self-renewal and maturation that are canonical to all stem cells. Epigenetic or developmental programs contribute to functional differences between cancer stem cells and non-cancer stem cells within a tumour clone that the model assumes would be genetically identical[13,16]. The cancer stem cell and clonal evolution models are the subject of intense debate and often considered to be mutually exclusive[16,17]. The cancer stem cell model focuses on the concept of functional heterogeneity but does not take into account tumour evolution, intratumoral genetic variation, or the existence of genetically distinct subclones. On the other hand, the clonal evolution model focuses on genetic heterogeneity without considering

[1]Division of Stem Cell and Developmental Biology, Campbell Family Institute for Cancer Research/Ontario Cancer Institute, Toronto, Ontario M5G 1L7, Canada. [2]Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5G 1L7, Canada. [3]Department of Pathology, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA. [4]Department of Medical Oncology and Hematology, Princess Margaret Hospital, and Department of Medicine, University of Toronto, Toronto, Ontario M5G 2M9, Canada. [5]Hartwell Center for Bioinformatics and Biotechnology, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA.
*These authors contributed equally to this work.

the functional variation that might exist intratumorally within individual genetic subclones. As a first step to resolve the basis for intratumoral heterogeneity, functional assays must be combined with genetic analysis to determine whether tumours contain genetically distinct subclones of tumour-initiating cells.

Philadelphia chromosome acute lymphoblastic leukaemia (Philadelphia-positive (Ph^+) ALL) is an ideal disease in which to study the relationship between intratumoral clonal diversity, genetic alterations and cellular growth properties because it is considered a single clinical entity with identifiable and recurrent genetic abnormalities. Detailed studies have revealed a number of genetic alterations, notably deletions of the lymphoid transcriptional regulator *IKAROS* (also called *IKZF1*), *PAX5*, *EBF1*, as well as deletions involving *CDKN2A/B* that cooperate with *BCR–ABL1* in lymphoid leukaemogenesis[18]. Furthermore, the association of *IKZF1* deletion in Ph-negative ALL with poor patient outcome predicts that it will be possible to link specific genetic alterations with function[19]. Here we report development of a robust Ph^+ ALL xenograft system that was used to carry out a combined genetic and functional study of the genetic diversity of functionally defined tumour-initiating cells derived from diagnostic patient samples.

Modelling human Ph^+ ALL in xenografts

To determine whether a single genetic subtype of leukaemia like Ph^+ ALL exhibits uniform growth properties, we used three xenograft models of increasing immune deficiency: NOD.CB17-*Prkdc*^{scid}/J (NOD/SCID) mice; NOD/SCID mice treated with anti-CD122 to deplete innate immune cells (NS122)[20,21]; or NOD/SCID mice with deletion of the common gamma (γ)-chain (NSG)[22] (Supplementary Fig. 1). Diagnosis samples from 18 of 20 Ph^+ ALL patients efficiently engrafted NS122 mice (Supplementary Fig. 2) and recapitulated numerous aspects of the human disease including tumour dissemination, immunophenotype (Supplementary Fig. 3) and morphology (Supplementary Fig. 4). However, 10 of 20 patient samples caused clinically manifest disease before 15 weeks and were categorized as aggressive group 1 samples, whereas the remaining xenograft mice appeared healthy until they were killed and these were classified as non-aggressive group 2 samples (Fig. 1a). Accordingly, the leukaemic burden in bone marrow and systemic dissemination was significantly higher in group 1 versus group 2 samples (Fig. 1b, c, Supplementary Fig. 2b and Supplementary Table 1). Notably, group 1 samples engrafted all recipient types (Fig. 1d and Supplementary Fig. 5) even when transplanted at near-limiting dose (Supplementary Fig. 6). By contrast, group 2 samples failed to engraft to NOD/SCID mice (Fig. 1d and Supplementary Fig. 5) despite injection of 50-fold more cells (Supplementary Fig. 7). Of two group 2 samples (patient 2-5 and 2-6) unable to engraft NS122 recipients, one engrafted NSG mice (Supplementary Fig. 8). The extent of leukaemic dissemination was similar in NS122 or NSG mice for both group 1 and group 2 samples, although NSG mice had higher peripheral engraftment levels (Supplementary Fig. 9).

Genetic basis of functional heterogeneity

To examine a possible genetic basis for the distinct xenograft growth properties between group 1 and group 2 samples, genome-wide CNA profiling was undertaken. Overall, the frequency of genetic alterations in *IKZF1* (84%), *CDKN2A/B* (50%) and *PAX5* (50%) of our 20 samples paralleled previous studies[18]. A similar frequency of group 1 and group 2 samples had focal and complete deletions of the *IKZF1* locus. However, there were marked differences in the proportion with deletions of *CDKN2A/B* (group 1, 90%; group 2, 0%; $P = 0.0001$) and *PAX5* (group 1, 60%; group 2, 10%; $P = 0.057$) genes (Fig. 2e and Supplementary Table 2). Genomic quantitative polymerase chain reaction (qPCR) confirmed that the *CDKN2A/B* locus was not deleted in representative group 2 samples (Supplementary Table 3) nor hypermethylated in an independent cohort of Ph^+ ALL patients (Supplementary Fig. 10).

Because previous studies have reported a positive association between the efficiency of xenograft engraftment and clinical outcome in acute myeloid leukaemia (AML) patients[23], we compared the survival of group 1 and group 2 samples. We found a trend towards poorer outcome of group 1 patients with increased early relapse, although significance was not reached owing to the small sample number (Fig. 2f, $P = 0.08$). Limiting dilution analyses (LDA) of 11 patient samples showed that the leukaemia-initiating-cell frequency in group 1 samples was 80-fold higher than group 2 samples (Fig. 2g). The leukaemia-initiating-cell frequency in one group 1 patient was 11%, similar to the leukaemia-initiating-cell frequency previously observed in comparable murine models[24]. Although comparison of absolute leukaemia-initiating-cell frequency in xenografts and syngeneic recipients is subject to some uncertainty[25], the relative difference between group 1 and group 2 samples is consistent with their distinct growth properties. Moreover, preliminary evidence indicates that leukaemic evolution, defined by increasingly aggressive and less restrictive xenograft growth upon serial passage, also correlates with reduced functional heterogeneity as reflected by increased leukaemia-initiating-cell frequency (Supplementary Fig. 11 and Supplementary Table 4). Collectively, we provide the first data linking engraftment properties and leukaemia-initiating-cell frequency to both specific genetic events in cancer and clinical outcome of patients.

Clonal dynamics of Ph^+ ALL pathogenesis

Despite widespread use of tumour xenografts, there are few studies comparing genetic alterations in primary samples versus xenografts[6]. To determine whether genetic abnormalities of the diagnostic sample are propagated upon transplantation, we tracked the clonal dynamics of ALL growth in xenografts by comparing CNA profiles of 12 diagnosis samples (eight group 1 and four group 2 tumours) with paired primary and secondary xenografts (Fig. 2a). Overall, xenografts did not select for a specific genotype as no single lesion was acquired by all tumours (Supplementary Fig. 12 and Supplementary Table 5). In six samples, five of which were group 1, xenografts exhibited the same distribution of CNA as the diagnostic patient sample, and detailed analysis of the antigen receptor (AgR) loci confirmed that the predominant clone present at diagnosis was propagated in xenografts (Supplementary Fig. 13).

By contrast, multiple xenografts derived from the six other patient samples (three group 1 and three group 2) harboured distinct genetic changes compared to the predominant diagnostic clone (Fig. 2b, c and Supplementary Fig. 14), while also sharing major CNA such as *IKZF1* and *CDKN2A/B* (data not shown). The presence of multiple xenograft recipients from the same patient sample with both identical and new CNA strongly indicates the existence of subclones, present at low levels in the diagnostic sample, that harbour additional genetic alterations. Although these CNA are newly emergent, we could detect them in other diagnostic samples, pointing to their clinical relevance (Supplementary Fig. 12). Our findings mirror recent CNA analysis of matched diagnosis and relapse ALL patient samples where the majority of relapsed cases represent the evolution of a new clone that is related to, but distinct from, the predominant diagnostic clone[11]. Notably, the clinical outcome of patients in whom the predominant clone at diagnosis was propagated in xenografts (CNA concordant) was different from patients where xenografts were engrafted with a minor subclone that outcompeted the predominant clone (CNA discordant). CNA concordant samples had poorer outcome even within our very small cohort (Fig. 2d). These data point to the need for larger validation studies and they indicate that genetic alterations, which lead to clonal predominance at diagnosis and competitive growth advantage in xenografts, are linked to poor survival.

To determine whether the detection of minor subclones might be hindered by out-competition in xenografts of dominant or aggressive clones, we analysed the CNA profiles of two patient samples transplanted at limiting (to engraft single leukaemia-initiating cells) and at non-limiting cell doses. In patient 1-8, non-limiting (bulk) cell doses

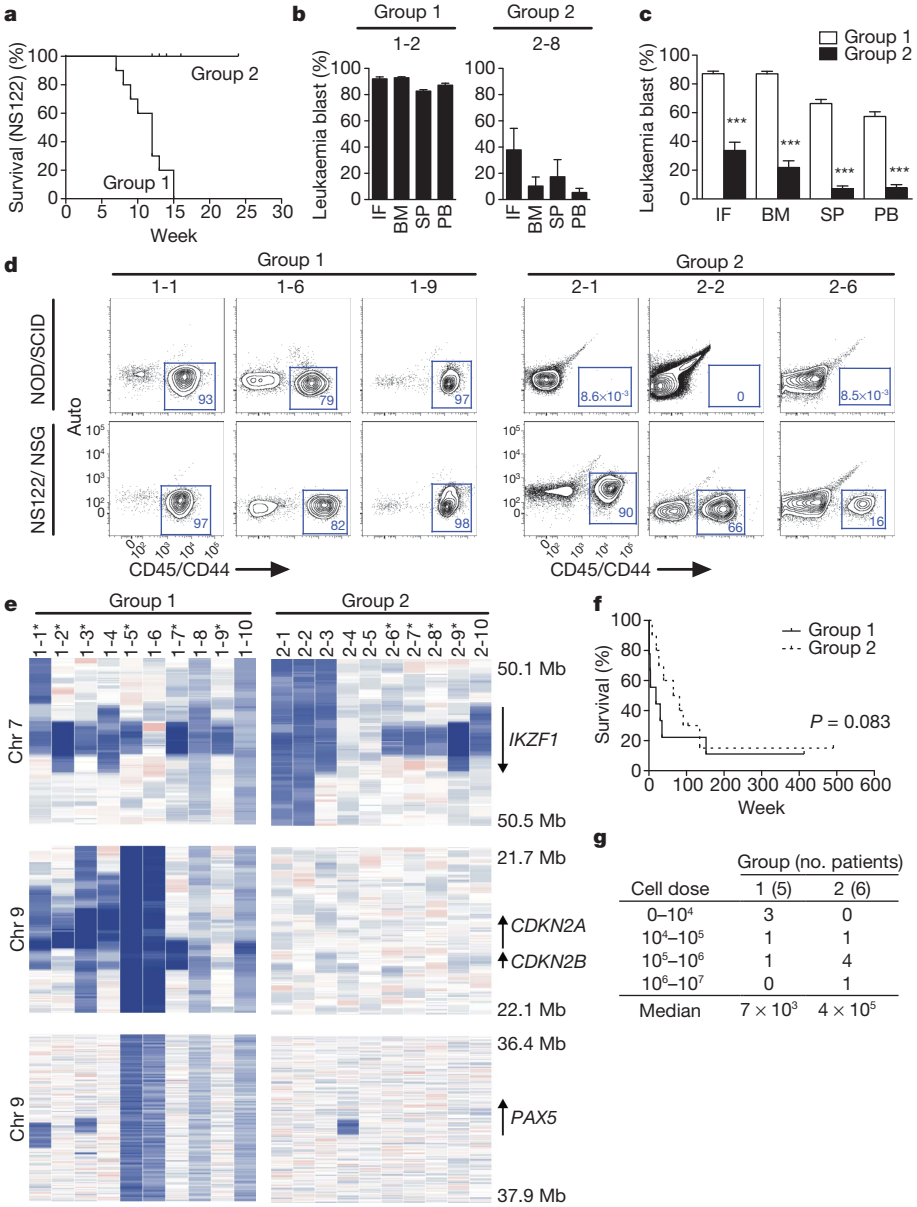


Figure 1 | Functional and genetic analysis of Ph$^+$ ALL. **a**, Survival of NS122 mice transplanted with 20 diagnostic Ph$^+$ ALL samples. Xenografts moribund before 15 weeks were categorized as group 1 (10 patient samples; $n = 90$) and the remainder that appeared healthy until they were killed for analysis were group 2 (10 patient samples; $n = 45$). **b**, Leukaemic burden in two representative patient samples, 1-2 and 2-8, determined by flow cytometry of the injected femur (IF), bone marrow (BM, left femur/tibiae), spleen (SP) and peripheral blood (PB) ($n = 4$ mice per sample; error bars, mean ± s.e.m.). **c**, Cumulative leukaemia engraftment in xenografts from panel **a** transplanted with group 1 and group 2 samples (error bars, mean ± s.e.m.; $^{***}P < 0.0001$). **d**, Comparison of leukaemic engraftment of group 1 ($n = 3$) and group 2 ($n = 3$) samples in NOD/SCID or NS122/NSG recipients. Human CD45 or CD44 (Supplementary Fig. 18) was used to evaluate chimaerism. **e**, CNA analysis of individual group 1 and group 2 patient samples using Affymetrix 6.0 SNP arrays. Regions containing *IKZF1* (top panel), *CDKN2A/B* (middle panel) and *PAX5* (bottom panel) deletions are shown. Asterisks indicate focal deletions in *IKZF1* involving exons 3–6, consistent with formation of dominant-negative *IKAROS* isoform IK6. Data are $\log_2$ ratio, median smoothing format (blue, deletion; white, normal; red, gain). See Supplementary Note for patient 1-8. **f**, Survival analysis of group 1 and group 2 patients ($P = 0.083$). **g**, Minimum cell dose required for leukaemia initiation in NS122 and NSG recipients from group 1 ($n = 5$) versus group 2 ($n = 6$) patient samples.

generated xenografts that differed at three major CNA compared to the patient sample (chromosome (Chr) 1 gain, Chr 8p deletion and Chr 8q gain) (Fig. 2e). Xenografts derived from limiting cell doses (clonal) differed from both non-limiting xenografts (lacked Chr 1 gain and 8p deletion) and the patient sample (Chr 8q gain) (Fig. 2e, m6). Patient sample 1-1 was highly aggressive in xenografts. All recipients of non-limiting cell doses and one from the limiting dose group contained the major diagnostic clone distinguished with a bi-allelic loss of *CDKN2A* (Fig. 2f, Chr 9). The other recipient from the limiting dose group did not have this CNA but retained the larger flanking CNA (Fig. 2f and Supplementary Fig. 15, m2). The topography of this lesion from each clone, together with the similarity of AgR regions in all recipients (Fig. 2e, f, top panels), indicates common genetic ancestry. Each clone remained stable after secondary transplantation (Supplementary Fig. 16). Thus, our data provide formal evidence that ALL is composed of genetically distinct subclones that are present in varying proportions at diagnosis and have differing functional capacity in xenografts. Furthermore, our analysis reveals the sequence of genetic events that probably occurred in these patients and provides unique insight into the evolution of clonal diversity during ALL leukaemogenesis.

Multi-clonal model of Ph$^+$ ALL pathogenesis

To gain an insight into the evolutionary processes that underlie the generation of dominant and minor subclones, we combined functional

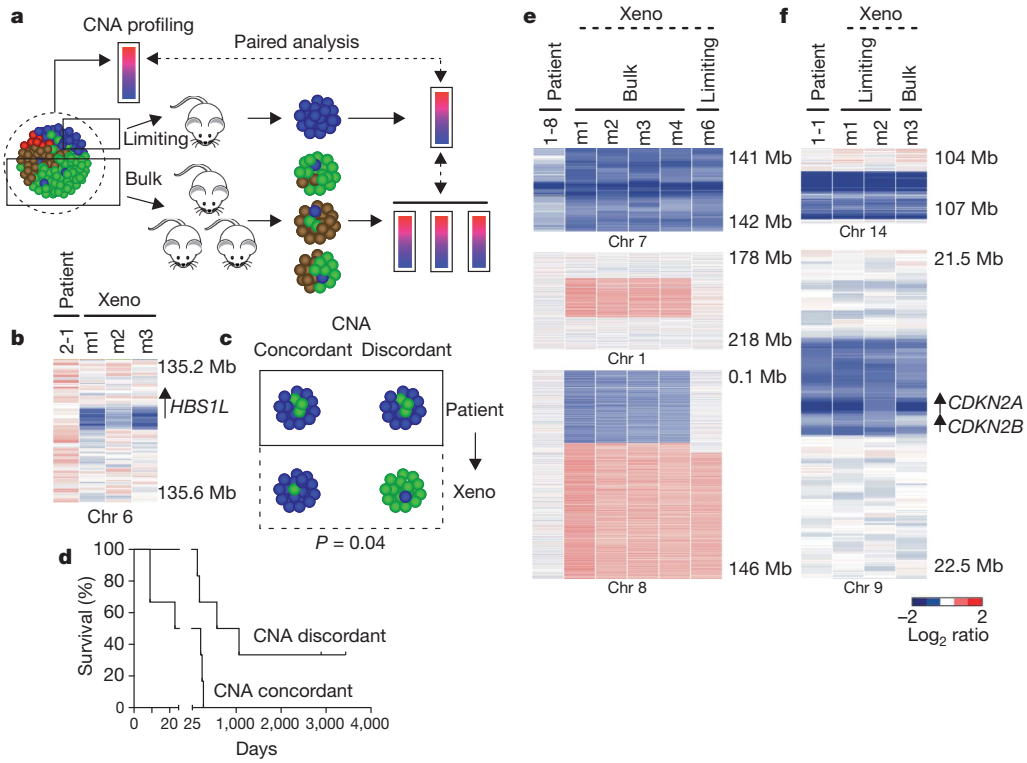


Figure 2 | Clonal dynamics of Ph$^+$ ALL upon transplant into xenografts. **a**, Schema of experimental design for tracking primary leukaemia clones in xenografts using limiting and bulk cell doses. **b**, Representative patient sample displaying a new CNA in each of three xenografts in the *HBS1L* locus that was below SNP array detection limit in the patient sample (additional examples on five patients in Supplementary Fig. 14). **c**, Pictorial representation of outgrowth of subclones (CNA discordant) or recapitulation of the predominant diagnostic clone (CNA concordant) after transplant into xenografts. **d**, Survival of patients that generated CNA concordant and discordant xenografts. **e**, **f**, CNA profiling of xenografts transplanted with limiting and bulk doses of patient cells. In patient 1-8 (**e**), multiple CNA present on Chr 1 (middle) and Chr 8 (bottom) were detected in all recipients (m1–m4) at bulk cell doses (1×10^6 per mouse) that were not detected in the diagnostic sample. At limiting cell dose, one of three mice was engrafted (5,000 cells per mouse, m6) and had partial similarity to the other bulk recipients (Chr 7 AgR deletion, top panel; Chr 8q gain, bottom panel) and the diagnostic sample (Chr 7, AgR deletion). In patient 1-1 (**f**), two of nine mice were engrafted at limiting cell doses (50 cells per mouse; m1, m2) and a single recipient at non-limiting cell dose ($\sim 1 \times 10^6$ cells; m3). The diagnostic sample, m1 and m3 share common CNA in *CDKN2A*, not present in m2 (bottom panel). All samples share common CNA at the AgR locus (top panel). Data are $\log_2$ ratio, median smoothing format (blue, deletion; white, normal; red, gain).

analysis obtained from xenografts with clonal analysis carried out by CNA profiling on multiple xenografts derived from a group 1 and group 2 sample. For both patient samples, the dominant diagnostic was not detected in xenografts, rather they were repopulated with several related but distinct genetic subclones. Detailed tracking of CNA provided an unprecedented opportunity to gain an insight into the sequence of lesion acquisition in independent subclones. For example, in patient 1-6, deletion of the AgR region of Chr 11 in all recipients indicates that this change was an early event in disease pathogenesis that was shared by all subclones that outgrew after transplant, followed by a gain in a region of Chr 6q present in leukaemic cells from two xenografts or a deletion at the Chr 14 immunoglobulin heavy chain (*IGH*) locus in a different subclone found in another xenograft (Fig. 3a). Patient 2-9 displayed a CNA (gain) in Chr Xp (Fig. 3b, Chr X) at diagnosis. All xenografts displayed varying loss of this region, indicating that this CNA was acquired early in pathogenesis of this patient sample. Only two recipients contained a subsequent Chr 9q deletion (m37 and m39), with further clonal divergence leading to a deletion in a region of Chr 8q (seen in m37) and duplication of Chr 8q (seen in m39) (Fig. 3b, Chr 8 and 9). Interestingly, xenograft m41 also displayed duplication in Chr 8q, but lacked a deletion in chromosome 9. Therefore, it remains unclear which CNA (Chr 9 deletion or Chr 8q duplication) was acquired first in the patient. Data are summarized pictorially for both patient samples (Fig. 3, right panel). These data indicate that multiple tumour clones coexist in the diagnostic patient sample, and that these clones undergo divergent evolution from the diagnostic clone, supporting the branching model of tumour progression[12].

Intratumoral heterogeneity may promote clonal evolution by increasing the number of selectable traits under any given stress. Therefore, genetic diversification is probably important for tumour survival. Recently, the degree of genetic diversification has been linked to clinical aggressiveness of breast tumours[26], and is associated with metastasis in pancreatic cancer[4,5]. To determine whether the various subclones we identified in patient sample 1-6 could evolve further, two primary xenografts were transplanted into secondary recipients. All CNA from primary xenografts were detected in secondary recipients, indicating overall stability of each subclone (Supplementary Fig. 17a). However, new CNA were detected in three of nine recipients, indicating that ongoing evolution and further progression of disease can occur, although it seems to be largely stochastic (Supplementary Fig. 17b). Although limiting dilution studies were not performed to completely rule out the possibility that additional minor subclones, undetected in the primary xenograft, contributed to these new CNA, these data indicate that genetic diversification can continue in the xenograft.

Discussion

Here we establish that individual Ph$^+$ ALL samples at diagnosis are composed of genetically diverse subclones that are related through a complex evolutionary process. These subclones vary in their xenograft growth properties and leukaemia-initiating-cell frequency. Furthermore, Ph$^+$ ALL patient samples can be segregated into two subgroups on the basis of functional xenograft growth properties and specific genetic mutations. The group that lost *CDKN2A/B* and had a tendency to poorer

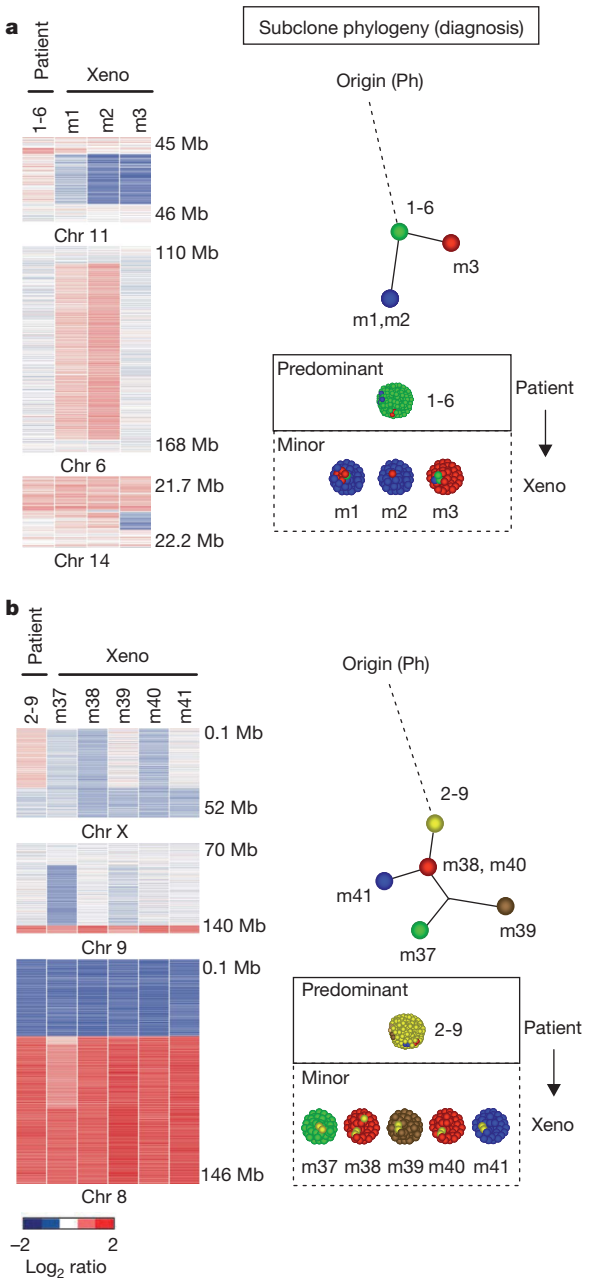


Figure 3 | Detection of genetically diverse leukaemia-initiating cells in Ph$^+$ ALL. **a**, CNA profiling shows that three engrafted recipients from sample 1-6 share a common CNA on Chr 11 (top) that is not detected in the diagnostic sample, but each is distinct for CNA on Chr 6 (middle) and Chr 14 (bottom). **b**, In sample 2-9, all five engrafted recipients shared a deletion (at varying degrees) in a region of Chr Xp (top), two recipients displayed a common CNA region of Chr 9q (middle; m37, m39), and three recipients had deletions (m37) and duplications (m39, 41) on regions of Chr 8p (bottom). A phylogenetic tree depicting the relationship between major and minor genetic subclones present in the diagnostic sample and a summary of the distinct xenograft subclones are shown to the right of each sample. The dashed line represents clonal evolution from disease origin (Ph, Philadelphia chromosome).

survival correlated with aggressive dissemination in xenografts and higher leukaemia-initiating-cell frequency. These results are consistent with the aggressive, tyrosine-kinase-inhibitor-resistant Ph$^+$ ALL seen in murine models where *Cdkn2a* is lost. The discovery that, at diagnosis, genetically distinct samples and subclones already possess variably aggressive growth properties points to the need to develop effective therapies to eradicate all intratumoral genetic subclones to prevent further evolution and recurrence. The ability to segregate even minor subclones in xenografts is a powerful tool for the preclinical development of new therapeutic strategies.

The segregation of individual genetic subclones in xenografts provided an opportunity to reconstruct the functional genetic ancestry of subclones present in diagnostic samples. Our data illustrate that leukaemic progression can occur in either a linear or branching fashion, with multiple genetic subclones evolving either in succession or in parallel, respectively. The xenograft growth characteristics of minor subclones were distinct: sometimes they out-competed the dominant clone, whereas in others (mostly group 1), they did not. The competitive advantage of the dominant clone also appeared to associate with poorer clinical outcome, although the size of the cohort was small. We speculate that some genetic events, such as loss of *CDKN2A/B*, contribute to clonal predominance at diagnosis and competitive xenograft growth. By contrast, the reduced competitive advantage of minor subclones indicates that additional genetic events are required for increased aggressiveness. However, if minor subclones survive therapy, further evolution and expansion could occur, leading to future relapse, consistent with previous genetic studies of paired diagnosis and relapse samples[11]. Because gene silencing and other epigenetic events contribute to tumour progression, our results also highlight the need for genome-wide methylation analysis of individual subclones. Evolution by branching increases subclonal complexity and underscores the importance of gaining a better molecular understanding of each subclone within a tumour.

Outgrowth of subclones in serial xenografts can only be sustained by leukaemia-initiating cells, and our findings establish that genetic diversity occurs in this functionally important cell type. Moreover, the discovery that specific genetic events influence leukaemia-initiating-cell frequency and that genetically distinct leukaemia-initiating cells evolve through a complex evolutionary process indicates that a close connection must exist between genetic and functional heterogeneity. This has several implications that may bring together the clonal evolution and cancer stem cell models. We remain cautious in extrapolating our findings in the absence of prospective isolation proving the existence of leukaemia stem cells in Ph$^+$ ALL; however, strong evidence is accumulating for the existence and relevance of leukaemia stem cells in other forms of leukaemia[27] and it is likely that genetically diverse leukaemia stem cells will eventually be found. If the leukaemia-initiating-cell/leukaemia stem cell link is established, we can speculate that leukaemia stem cells are not static but are able to evolve genetically and represent units of selection in tumour evolution. As tumours evolve, the frequency of leukaemia stem cells increases, eventually progressing to a highly advanced state that might no longer adhere to a cancer stem cell model. The high leukaemia-initiating-cell frequency that we observed in some samples and from several murine models[16,24] supports this idea. Finally, as tumours are composed of genetically diverse subclones, prospective isolation of leukaemia stem cells/cancer stem cells needs to be interpreted with considerable care, as fractionation of cancer stem cell and non-cancer stem cell populations could segregate genetically distinct subclones with variable tumour-initiating-cell capacity as opposed to genetically identical cells with differing epigenetic/developmental programs. Because the hierarchy model posits that cancer stem cells give rise to non-cancer stem cells, future studies must account for subclonal diversity and establish the genetic identity of cancer stem cells and non-cancer stem cells. Overall, our findings indicate that there may be more commonalities between clonal evolution and cancer stem cell models of cancer than previously thought and that future studies may lead to a unification of these concepts.

METHODS SUMMARY

Diagnostic Ph$^+$ ALL patient samples were intrafemorally transplanted into female NOD.CB17-*Prkdc*^{scid}/J (NOD/SCID) mice, NOD/SCID mice treated with mouse anti-CD122 (as previously described) and NOD.Cg-*Prkdc*^{scid} *Il2rg*^{tm1Wjl}/SzJ (NSG) mice. Xenograft recipients were monitored for disease sickness, and chimaerism was evaluated in various haematopoietic tissues using flow cytometry. DNA copy

number alteration (CNA) was carried out with Affymetrix 6.0 SNP arrays on the diagnostic patient sample and corresponding xenografts.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 10 June; accepted 3 December 2010.

1. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100,** 57–70 (2000).
2. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194,** 23–28 (1976).
3. Barrett, M. T. *et al.* Evolution of neoplastic cell lineages in Barrett oesophagus. *Nature Genet.* **22,** 106–109 (1999).
4. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467,** 1109–1113 (2010).
5. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467,** 1114–1117 (2010).
6. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464,** 999–1005 (2010).
7. Bateman, C. M. *et al.* Acquisition of genome-wide copy number alterations in monozygotic twins with acute lymphoblastic leukemia. *Blood* **115,** 3553–3558 (2010).
8. Hong, D. *et al.* Initiating and cancer-propagating cells in TEL-AML1-associated childhood leukemia. *Science* **319,** 336–339 (2008).
9. Li, A. *et al.* Sequence analysis of clonal immunoglobulin and T-cell receptor gene rearrangements in children with acute lymphoblastic leukemia at diagnosis and at relapse: implications for pathogenesis and for the clinical utility of PCR-based methods of minimal residual disease detection. *Blood* **102,** 4520–4526 (2003).
10. Zuna, J. *et al.* TEL deletion analysis supports a novel view of relapse in childhood acute lymphoblastic leukemia. *Clin. Cancer Res.* **10,** 5355–5360 (2004).
11. Mullighan, C. G. *et al.* Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science* **322,** 1377–1380 (2008).
12. Greaves, M. Cancer stem cells: back to Darwin? *Semin. Cancer Biol.* **20,** 65–70 (2010).
13. Dick, J. E. Stem cell concepts renew cancer research. *Blood* **112,** 4793–4807 (2008).
14. Bruce, W. R. & Van Der Gaag, H. A quantitative assay for the number of murine lymphoma cells capable of proliferation *in vivo*. *Nature* **199,** 79–80 (1963).
15. Diehn, M., Cho, R. W. & Clarke, M. F. Therapeutic implications of the cancer stem cell hypothesis. *Semin. Radiat. Oncol.* **19,** 78–86 (2009).
16. Shackleton, M., Quintana, E., Fearon, E. R. & Morrison, S. J. Heterogeneity in cancer: cancer stem cells versus clonal evolution. *Cell* **138,** 822–829 (2009).
17. Marusyk, A. & Polyak, K. Tumor heterogeneity: causes and consequences. *Biochim. Biophys. Acta* **1805,** 105–117 (2010).
18. Mullighan, C. G. *et al.* BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. *Nature* **453,** 110–114 (2008).
19. Mullighan, C. G. *et al.* Deletion of *IKZF1* and prognosis in acute lymphoblastic leukemia. *N. Engl. J. Med.* **360,** 470–480 (2009).
20. McKenzie, J. L., Gan, O. I., Doedens, M. & Dick, J. E. Human short-term repopulating stem cells are efficiently detected following intrafemoral transplantation into NOD/SCID recipients depleted of $CD122^+$ cells. *Blood* **106,** 1259–1261 (2005).
21. Taussig, D. C. *et al.* Anti-CD38 antibody-mediated clearance of human repopulating cells masks the heterogeneity of leukemia-initiating cells. *Blood* **112,** 568–575 (2008).
22. Shultz, L. D. *et al.* Human lymphoid and myeloid cell development in NOD/LtSz-*scid IL2R*γ^{null} mice engrafted with mobilized human hemopoietic stem cells. *J. Immunol.* **174,** 6477–6489 (2005).
23. Pearce, D. J. *et al.* AML engraftment in the NOD/SCID assay reflects the outcome of AML: implications for our understanding of the heterogeneity of AML. *Blood* **107,** 1166–1173 (2006).
24. Williams, R. T., den Besten, W. & Sherr, C. J. Cytokine-dependent imatinib resistance in mouse $BCR\text{-}ABL^+$, Arf-null lymphoblastic leukemia. *Genes Dev.* **21,** 2283–2287 (2007).
25. Dick, J. E. Looking ahead in cancer stem cell research. *Nature Biotechnol.* **27,** 44–46 (2009).
26. Park, S. Y., Gonen, M., Kim, H. J., Michor, F. & Polyak, K. Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *J. Clin. Invest.* **120,** 636–644 (2010).
27. Tehranchi, R. *et al.* Persistent malignant stem cells in del(5q) myelodysplasia in remission. *N. Engl. J. Med.* **363,** 1025–1037 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to thank S. Minkin for statistical analysis of patient outcome, the Dick Laboratory and B. Neel for critical review of this manuscript, M. Cooper for editorial assistance, and P. A. Penttilä, L. Jamieson, J. Yuan and S. Zhao for preparative flow cytometry. This work was supported by funds from Canadian Institutes for Health Research (CIHR) studentships (F.N., S.D.), the Pew Charitable Trusts (C.G.M.), The Stem Cell Network of Canadian National Centres of Excellence, the Canadian Cancer Society and the Terry Fox Foundation, Genome Canada through the Ontario Genomics Institute, Ontario Institute for Cancer Research with funds from the province of Ontario, the Leukemia and Lymphoma Society, the Canadian Institutes for Health Research, a Canada Research Chair, and the American and Lebanese Syrian Associated Charities of St Jude Children's Research Hospital. This research was funded in part by the Ontario Ministry of Health and Long Term Care (OMOHLTC). The views expressed do not necessarily reflect those of the OMOHLTC.

Author Contributions F.N. designed study, analysed data and prepared figures. F.N., C.G.M., J.C.Y.W., A.P., S.D. and L.A.P. performed experiments. M.D.M. provided patient samples. J.C.Y.W. and M.D.M. provided patient outcome data. J.M. performed paired and unpaired segmentation analysis of SNP array data. F.N. and C.G.M analysed and interpreted SNP data. C.G.M., J.C.Y.W., S.D. and J.R.D. critically reviewed and edited the manuscript. F.N. and J.E.D. wrote the manuscript. J.E.D. supervised the study.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.E.D. (jdick@uhnres.utoronto.ca).

METHODS

Patient samples. Patient samples (primarily peripheral blood) were obtained from newly diagnosed Philadelphia-positive acute lymphoblastic leukaemia patients according to pre-established guidelines approved by the Research Ethics Board of University Health Network. Three out of twenty patient samples were donated by the Quebec Leukaemia Cell Bank. All samples were frozen viably using standard protocols in FCS + 10% DMSO and stored long term at −150 °C. Cell viability, as assessed immediately after thawing, was greater then 90% for all cases. Detailed patient data are provided in Supplementary Tables 6 and 7.

Xenotransplantation assay and analysis. NOD.CB17-*Prkdc*^{scid}/J (NOD/SCID) and NOD.Cg-*Prkdc*^{scid}*Il2rg*^{tm1Wjl}/SzJ (NSG) mice were bred according to protocols established and approved by the Animal Care Committee at University Health Network. Ten-to-twelve-week-old old mice were sublethally irradiated at 225 cGy 24 h before transplant. NOD/SCID mice were also treated with mouse anti-CD122 monoclonal antibody (NS122) after sublethal conditioning as previously described[20]. Only female mice were used in these studies[28]. NSG mice were periodically genotyped using standard PCR protocols (wild-type common forward primer, 5′-GTGGGTAGCCAGCTCTTCAG-3′; wild-type reverse, 5′-CCTGGAGC TGGACAACAAAT-3′; mutated reverse, 5′-GCCAGAGGCCACTTGTGTAG-3′). All primer sets were obtained from Jackson Laboratories website.

Frozen leukaemia cells were thawed by drop-wise addition of IMDM + 0.2 mg l^{-1} DNase (Roche Applied Science). After centrifugation, cells were resuspended in IMDM and counted using ViCell XR (Beckman coulter) or by trypan blue exclusion. For transplantation, cells were placed in microfuge tubes and spun down to remove excess media. Cells were resuspended in the correct volume (25 µl per mouse) for transplant using IMDM + 1% FCS. Intrafemoral transplant was performed as previously described[29]. Briefly, mice were anaesthetized using isoflurane. The right knee of mice was bent and drilled with a 27.5-g needle and followed by injection of cells with a 28.5-g insulin syringe (BD Biosciences).

After transplant, animals were monitored for the appearance of disease symptoms, such as weight loss, hunch-back, decreased activity, and killed soon after. Mice transplanted with samples that did not induce disease symptoms (in NS122 or NSG recipients) were killed by 16–24 weeks. Upon death, injected right femur (IF), non-injected bones (left femur, right and left tibiae, bone marrow (BM)), spleen and peripheral blood were removed and analysed for the presence of human leukaemia blasts using flow cytometry. Aliquots of cells were stained in 96-well round-bottomed plates (BD Falcon). Human cells were distinguished from mouse cells using human-specific CD45PC7 and CD44PE (Supplementary Fig. 18). B-cell-specific markers (various combinations of IgM FITC, CD19 PC5 (Beckman Coulter), CD20 APC7, CD10 APC, CD34 APC7) were used to evaluate blast phenotype after transplant. Detection of normal haematopoietic stem cell activity was monitored using CD33 PC5 (Beckman Coulter) or APC. T-cell engraftment in NSG mice was monitored using human CD3 FITC or APC. All fluorochromes were obtained from BD Biosciences unless otherwise indicated. Mice were considered to be engrafted when multiple human leukaemia markers (for example, $CD44^+CD45^+CD34^+CD19^+$) were detected by at least the 0.5% threshold. Virtually all cases of engraftment were well above this threshold. All flow cytometry analysis was performed on the LSRII (BD Biosciences). The remaining marrow (IF + BM) was frozen viably in FCS + 10% DMSO.

DNA SNP microarray analysis. DNA isolated from patients at diagnosis and xenograft samples was analysed using Affymetrix 6.0 SNP arrays. SNP array data were analysed using dChip no. 4 software using a reference algorithm and circular binary segmentation as previously described[11,18,19,30–32]. To distinguish inherited from somatic DNA copy number alterations for primary patient samples lacking matched normal DNA, putative variants identified by unpaired segmentation were filtered using public copy number polymorphism databases[33,34], and an in-house database of SNP array data from several hundred samples. SNP array data are available at dbGaP (phs000329.v1.p1) and is also hosted through St Jude Children's Research hospital (http://hospital.stjude.org/forms/genome-download/request/). Sample information is shown in Supplementary Table 7.

Quantitative PCR of the *CDKN2A/B* locus. Primers for genomic quantitative PCR were designed using Primer Express 3.0 (Applied Biosystems), and are listed in Supplementary Table 3. Taqman RNase P primers (Applied Biosystems) were used for control amplification. Fifty nanograms of leukaemic blast DNA or control human DNA was amplified using a 7500 Real-Time PCR system and 7500 System Software (Applied Biosystems), using the 7500 universal cycling conditions: 50 °C for 2 min, followed by 95 °C for 10 min, then 40 cycles of 95 °C for 1 min and 60 °C for 1 min. Standard curves for each *CDKN2A/B* exon and RNase P were generated using normal human DNA. Assays were performed in duplicate. *CDKN2A/B* exon-specific copy number values were normalized by dividing the value obtained for each test reaction by the paired value obtained for RNase P for each sample. Cutoffs of 0.65 and 0.3 were used to identify hemizygous and homozygous PAX5 deletions, respectively.

Statistical analysis. All data were analysed using GraphPad Prism version 5.00 for Mac OS X (http://www.graphpad.com). The Mann–Whitney *U*-test was used to assess statistically significant differences in chimaerism in xenografts. Clinical outcome data was analysed using the Gehan–Wilcoxon method based on random 10,000 permutations. Because the normal distribution of data can only be assumed in larger cohorts, this modified test more accurately computes *P*-values for small patient cohorts.

28. Notta, F., Doulatov, S. & Dick, J. E. Engraftment of human hematopoietic stem cells is more efficient in female NOD/SCID/IL-2Rgc-null recipients. *Blood* **115,** 3704–3707 (2010).
29. Mazurier, F., Doedens, M., Gan, O. I. & Dick, J. E. Rapid myeloerythroid repopulation after intrafemoral transplantation of NOD-SCID mice reveals a new class of human stem cells. *Nature Med.* **9,** 959–963 (2003).
30. Lin, M. *et al.* dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics* **20,** 1233–1240 (2004).
31. Mullighan, C. G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446,** 758–764 (2007).
32. Pounds, S. *et al.* Reference alignment of SNP microarray signals for copy number analysis of tumors. *Bioinformatics* **25,** 315–321 (2009).
33. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genet.* **36,** 949–951 (2004).
34. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genet.* **40,** 1166–1174 (2008).

ARTICLE

doi:10.1038/nature09652

Nascent transcript sequencing visualizes transcription at nucleotide resolution

L. Stirling Churchman[1] & Jonathan S. Weissman[1]

Recent studies of transcription have revealed a level of complexity not previously appreciated even a few years ago, both in the intricate use of post-initiation control and the mass production of rapidly degraded transcripts. Dissection of these pathways requires strategies for precisely following transcripts as they are being produced. Here we present an approach (native elongating transcript sequencing, NET-seq), based on deep sequencing of 3′ ends of nascent transcripts associated with RNA polymerase, to monitor transcription at nucleotide resolution. Application of NET-seq in *Saccharomyces cerevisiae* reveals that although promoters are generally capable of divergent transcription, the Rpd3S deacetylation complex enforces strong directionality to most promoters by suppressing antisense transcript initiation. Our studies also reveal pervasive polymerase pausing and backtracking throughout the body of transcripts. Average pause density shows prominent peaks at each of the first four nucleosomes, with the peak location occurring in good agreement with *in vitro* biophysical measurements. Thus, nucleosome-induced pausing represents a major barrier to transcriptional elongation *in vivo*.

Accumulating evidence now reveals that transcription elongation is not a straightforward read-out of the downstream DNA sequence. Co-transcriptional processing events dictate the covalent nature and fate of RNA transcripts[1]. Indeed many transcripts are targeted co-transcriptionally for rapid degradation and hence are effectively invisible to approaches that monitor mature messenger RNAs[2–4]. In addition to these processing events, the strong propensity of RNA polymerase (RNAP) to pause creates barriers to elongation and provides an opportunity for regulation and coordination of co-transcriptional events[5,6]. *In vitro*, RNAP pausing is found to be ubiquitous[7]. Biophysical approaches have provided a structural and energetic understanding of RNAP pausing which results from both intrinsic properties of the polymerase itself as well as interactions with its DNA template, including the presence of bound proteins (for example, histones)[8–12]. In the cell, elongation factors probably alter the energetic landscape of transcription, but the extent and mechanism of RNAP pausing in eukaryotic cells remain largely unknown. Bridging the divide between *in vivo* and *in vitro* transcriptional views requires approaches that visualize transcription with comparable precision afforded by *in vitro* transcriptional assays. More generally, the ability to monitor quantitatively nascent transcripts would provide broad insights into the roles and regulation of transcription initiation, elongation and termination in gene expression.

Historically, two strategies have been used to provide snapshots of transcriptional activity *in vivo*. In the first approach, RNAP is crosslinked to DNA and RNAP-bound DNA elements are identified by microarrays or deep sequencing[13,14]. Although providing a global view of RNAP binding sites, these measurements are of limited spatial and temporal resolution and do not reveal the identity of the transcribed strand or even whether RNAP molecules are engaged in transcription. In the second approach, transcription is halted *in vivo* and then reinitiated in isolated nuclei under conditions that allow labelling of nascent chains, thereby enabling them to be distinguished from bulk RNA[15,16]. Such 'nuclear run-on' strategies reveal actively transcribed DNA regions but require extensive manipulations that limit resolution and depend on the efficient re-initiation of transcription under non-physiological conditions.

To monitor the transcriptional states of unperturbed cells, we sought to determine the precise *in vivo* position of all active RNAP complexes. Here we present an approach (native elongating transcript sequencing, NET-seq) that accomplishes this goal by exploiting the extraordinary stability of the DNA–RNA–RNAP ternary complex[17] to capture nascent transcripts directly from live cells without crosslinking. The identity and abundance of the 3′ end of purified transcripts are revealed by deep sequencing[18], thus providing a quantitative measure of RNAP density with single nucleotide precision. Using NET-seq, we expose rapidly degraded transcription products, locate the position of RNAP pauses, and identify factors and chromatin structure that regulate these transcription events.

Quantifying transcription at nucleotide resolution

We focused on the transcription by RNAPII of protein-coding genes in the budding yeast *Saccharomyces cerevisiae*, although the NET-seq approach should be readily adaptable to other systems. To facilitate purification, we worked with a strain that endogenously expressed a functional variant of RNAPII with a 3×-Flag epitope attached to its third subunit (Rpb3). Log-phase cultures were collected by filtration and flash frozen in liquid nitrogen (Fig. 1a). After cryogenic lysis, RNAPII was efficiently immunoprecipitated (Supplementary Fig. 1a). We prepared the co-purified RNA for deep sequencing using a protocol that allows efficient RNA capture while minimizing bias[19], and sequenced 40 bases from the 3′ end. The alignment of these sequences to the yeast genome identified the final nucleotide that was incorporated by RNAPII, and the number of sequencing reads at each position along the genome indicated the density of transcriptionally active RNA polymerases at that site (Fig. 1b, alignment statistics displayed in Supplementary Table 1). A metagene analysis of RNAPII distribution across transcription units shows higher RNAPII density for the first 700 base pairs (bp) from the 5′ end (Fig. 1c), consistent with lower resolution observations seen using a global run-on approach[16].

Several observations indicate that we are detecting nascent transcription. First, we robustly capture transcripts from introns and regions

[1]Department of Cellular and Molecular Pharmacology, Howard Hughes Medical Institute, University of California, San Francisco and California Institute for Quantitative Biosciences, San Francisco, California 94158, USA.

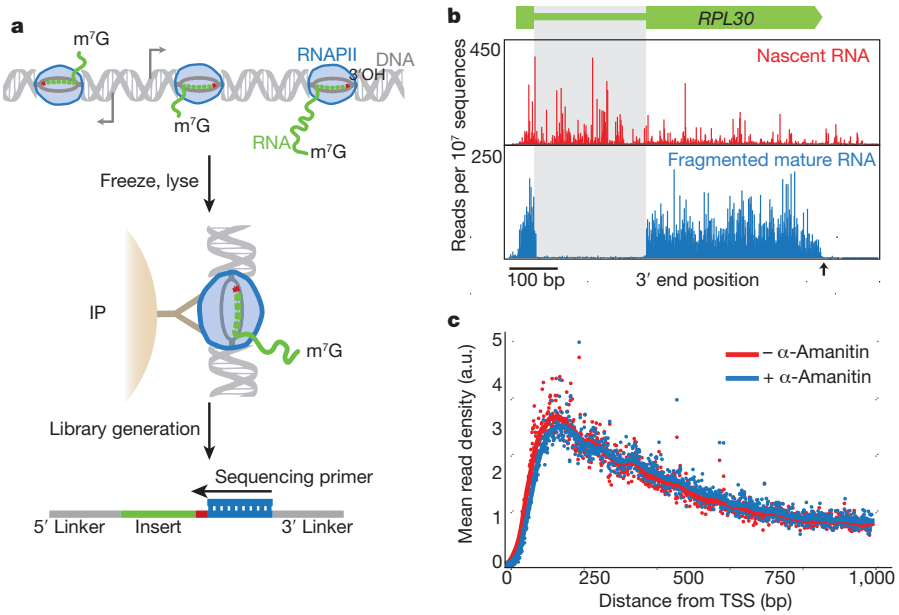


Figure 1 | NET-seq visualizes active transcription via capture of 3′ RNA termini. **a**, Schematic diagram of NET-seq protocol. A yeast culture is flash frozen and cryogenically lysed. Nascent RNA is co-purified via an immunoprecipitation (IP) of the RNAPII elongation complex. Conversion of RNA into DNA results in a DNA library with the RNA as an insert between DNA sequencing linkers. The sequencing primer is positioned such that the 3′ end of the insert is sequenced. m^{7}G refers to the 7-methylguanosine cap structure at the 5′ end of nascent transcripts. **b**, The 3′ end of each sequence is mapped to the yeast genome and the number of reads at each nucleotide is plotted at the *RPL30* locus for nascent RNA and lightly fragmented mature RNA. Note that for the nascent transcripts, the introns (grey box) and regions after the polyadenylation site (black arrow) are readily detected. **c**, Metagene analysis for well-expressed genes ($n = 471$, >1.5 reads per bp in both conditions) of the mean read density (arbitrary units, a.u.) in the presence and absence of transcription inhibitor, α-amanitin. TSS, transcription start site.

after polyadenylation sites; areas that are present in nascent transcripts but absent from mature messenger RNAs (Fig. 1b). Second, we verified that transcripts do not associate with RNAPII after cell lysis (Supplementary Table 2). Third, we saw negligible degradation of RNA under the immunoprecipitation conditions. Nevertheless, our library generation protocols prevent detection of co-purified degradation products by requiring that input RNAs have 3′ hydroxyl termini, as hydrolysis and degradation products primarily have terminal phosphates[20]. Finally, we saw that transcription did not proceed during processing of lysates as addition of the transcription inhibitor α-amanitin to the lysis buffer did not change the RNAPII density (Fig. 1c).

In addition to nascent transcripts, the RNAPII immunoprecipitation captures splicing intermediates (that is, the 5′ exon and the excised lariat). Their 3′ hydroxyl termini allow them to appear in our data at the 3′ ends of exons and introns (Supplementary Fig. 5). These observations indicate the widespread existence of co-transcriptional splicing in yeast and establish NET-seq as a powerful tool for studying such events.

Direct observation of transcription of unstable RNA

NET-seq monitors transcripts regardless of their stability, making it ideally suited to the analysis of unstable transcripts. Recent studies have revealed a class of cryptic unstable transcripts (CUTs) that are short (less than ~700 nucleotides), upstream and antisense to an annotated gene and rapidly degraded by the exosome[2–4,15,21]. Divergent transcription, yielding the production of antisense CUTs and mature messenger RNAs from the sense direction, is seen at many promoters in both yeast and metazoans. The observation of widespread divergent transcription was surprising and it remains unclear how antisense transcripts initiate and what biological function they may have. It is likely that the nucleosome-free region associated with promoters facilitates antisense transcription. Additionally, it has been suggested that antisense and sense transcription levels are co-dependent[15,21], as transcription in the sense direction could promote upstream antisense transcription (and vice versa) by creating negatively supercoiled DNA and recruiting factors that set permissive histone marks[22]. Critical evaluation of these hypotheses has been limited by the difficulty in quantitatively monitoring the levels of unstable antisense transcripts.

As NET-seq directly monitors the production of transcripts, we were able to quantify the relative amounts of nascent sense and antisense transcripts (Fig. 2a, b). We focused our analysis on promoters between genes encoded on the same strand (tandem genes), because in those instances, antisense transcripts can be clearly differentiated from the stable upstream transcript. To quantify divergent transcription, we integrated the transcript levels for the first 500 bp of transcribed DNA in each direction. Although we clearly observed divergent promoters, the large majority of promoters had much less antisense transcription than sense transcription; for more than half of the promoters, sense transcription was at least eight times higher than antisense transcription, and for 80% of the promoters the sense-to-antisense transcription ratio exceeded threefold (Fig. 2b). Notably, a comparison between the levels of sense and antisense transcription showed only modest correlation (Spearman correlation coefficient, $r_s = 0.34$) (Fig. 2c).

The above analysis establishes that antisense transcription is not an obligatory consequence of having an active promoter. What then dictates whether a promoter is directional? Transcription initiation is known to occur in nucleosome-free regions; however, we failed to see a correlation when we compared antisense transcription levels with published data[23] reporting on promoter nucleosome-free-region size and promoter average nucleosome occupancy (Supplementary Fig. 2a, b). We also investigated whether histone modifications associated with active promoters correlated with antisense transcription, as it was observed that H3 acetylation peaks in regions of antisense transcription in human fibroblasts[15]. Notably, we found a strong positive correlation ($r_s = 0.65$) between antisense transcription levels and earlier measurements of the levels of H4 (and to a lesser extent H3) acetylation enrichment[24] (Fig. 2d and Supplementary Fig. 2c, d).

Rpd3S promotes promoter directionality

The strong correlation between antisense transcription and H4 acetylation indicates that H4 acetylation may have a causative role in

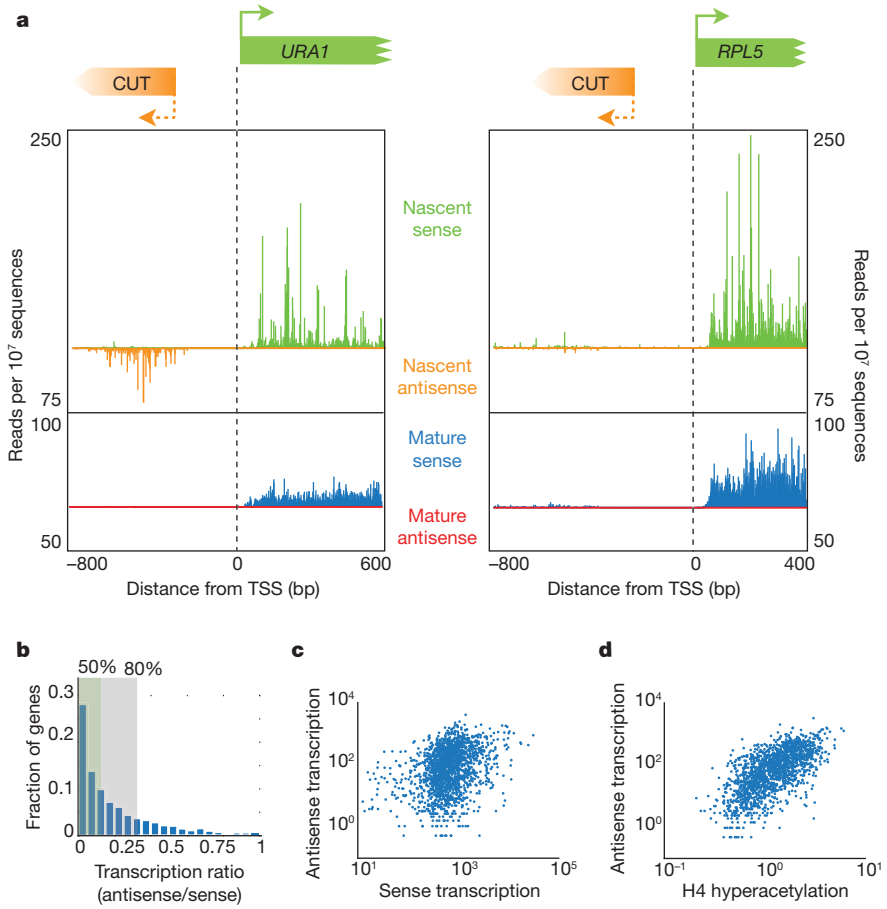


Figure 2 | Observation of divergent transcripts reveals strong directionality at most promoters. **a**, Nascent and mature transcripts initiating from *URA1* and *RPL5* promoters in the sense and antisense directions. Note that there are cryptic unstable transcripts (CUTs) in the antisense direction for *URA1* but not *RPL5*. **b**, A histogram of the transcription ratio (antisense/sense transcription levels) for 1,875 genes. The green and grey boxes indicate the subset of genes with a ratio of less than 1:8 and less than 1:3, respectively. **c**, Antisense transcription levels are plotted versus sense transcription for each tandem gene (Spearman correlation coefficient, $r_s = 0.34$). **d**, The level of antisense transcription for each promoter is plotted versus the local enrichment for H4 hyperacetylation using available data[24] ($r_s = 0.65$).

facilitating antisense transcription. To test this, we examined the effect on antisense transcription of loss of *RCO1*, a required and dedicated subunit of the Rpd3 small (Rpd3S) H4 deacetylation complex[25,26]. We focused on Rpd3S, as earlier studies had shown that it contributes to deacetylation of H4 in the 3′ region of transcripts and the large majority of antisense transcripts overlap the 3′ ends of upstream genes. Previous global studies of Rpd3S monitored accumulation of mature stable RNAs and so would not detect the effects of Rco1 on transient RNA species[25,26]. Our analysis revealed a pervasive increase (average fourfold) in unstable antisense transcription (Fig. 3a, b). This effect was the dominant transcriptional phenotype that we observed and was specific to antisense transcription: we found no systematic increase in RNAPII density at the beginning of sense transcripts (Supplementary Fig. 3). Importantly, antisense transcripts seen in the *rco1Δ* strain have the same transcription start sites and the same lengths as the wild-type transcripts, indicating that Rco1 is acting at the initiation stage of antisense transcription and does not affect termination (Fig. 3c). Additionally, we observed that deletion of *EAF3*, another subunit of Rpd3S, mimicked the increases seen in the *rco1Δ* data ($r_s = 0.88$, Supplementary Fig. 4). Thus, the primary function of the Rpd3S histone deacetylase complex seems to be to enforce promoter directionality.

This raises the question of how Rpd3S is recruited to positions designated for suppression of antisense transcription. The Rco1 and Eaf3 components of the Rpd3S complex bind H3 lysine 36 methylation marks made by Set2 and that binding activates the deacetylase activity of Rpd3S (refs 25–27). However, a distinct RNAPII-associated methyltransferase, Set1, has also been implicated in Rpd3S-dependent repression[28]. Moreover, even in the absence of methylation, RNAPII is capable of recruiting Rpd3S to gene bodies during transcription[29].

To investigate how Rpd3S is localized to suppress antisense transcription, we analysed nascent transcripts in cells lacking Set1 or Set2. *SET1* deletion caused a weak increase in antisense transcription in a manner that correlated only modestly with the *rco1Δ* and *eaf3Δ* data ($r_s = 0.36$ and $r_s = 0.38$ respectively; Supplementary Fig. 5). In contrast, deletion of *SET2* led to a pronounced increased in antisense transcription that was highly correlated with the *rco1Δ* and *eaf3Δ* data ($r_s = 0.88$ and $r_s = 0.89$ respectively; Supplementary Fig. 5). These data together with earlier work on the Set2/Rpd3S pathway indicate that the major mechanism for Rpd3S action on antisense transcription involves Set2 recruitment to elongating RNAPII via Ser 2 phosphorylation on its carboxy-terminal domain[30]. This in turn, through the Set2 methylation activity, allows recruitment of Rpd3S to the 3′ ends of genes, suppressing antisense transcription from downstream nucleosome-free regions. Future challenges will be to explain how histone acetylation in the body of antisense transcripts can affect transcription initiation, and to determine other mechanisms that localize Rpd3S, particularly for the handful of antisense transcripts that do not overlap the 3′ ends of genes.

Pausing occurs throughout transcription elongation

The ability of NET-seq to map the density of nascent transcripts enables in-depth investigation of the extent and sources of RNAP pausing *in vivo*. Our data revealed strong and highly reproducible spikes in the density of 3′ ends of nascent transcripts along a given gene indicative of RNAPII pause sites (for example, *GPM1*; Fig. 4a). We developed an algorithm to identify RNAPII pause positions that finds points where the read density is at least three standard deviations above the mean in a local 200-bp window. We found that pauses occur frequently throughout the body of RNA messages and are evenly distributed after the first ~700 bp (Fig. 4b and Supplementary Fig. 6). The high density of pauses was not an artefact of library generation and sequencing biases, as we detected tenfold fewer spikes in data from messenger RNA lightly fragmented by alkaline hydrolysis (Supplementary Fig. 7). Notably, 70% of the more than 2×10^5 pause sites that we identified had an A at the 3′ end of the transcript.

Additionally, there was a preference for the pause to be followed immediately by a T and then G (Fig. 4c). None of these biases was seen in the control sample of fragmented mRNA (Supplementary Fig. 7a).

Largely from *in vitro* studies, one mechanism of RNAP pausing has been shown to involve backtracking: after encountering a blockage, RNAP reverses direction and moves upstream[31]. In the backtracked state, the 3′ end of the RNA transcript is no longer aligned with the active site and RNAP must either return to the initial pause site or cleave the transcript. The latter option is aided by the presence of the elongation factor TFIIS (Dst1 in yeast) that enhances RNAP's intrinsic RNA cleavage activity (Fig. 5a)[32,33]. Although the role of TFIIS is well established *in vitro*, its mechanism *in vivo* has been less explored[34–36].

To investigate the role that backtracking has in pausing *in vivo*, we deleted *DST1* and repeated the NET-seq assay. Notably, we saw a large-scale downstream shift in the position of the pauses, an average of 5–18 bp (Fig. 5b, c). This shift was observed for ~75% of the pauses (Supplementary Fig. 8) and was accompanied by a global change in the sequences surrounding pause sites; the preference for A at the pause was lost and instead there was a strong preference for T immediately downstream of the pause (Fig. 5d). These observations confirm that the observed spikes in NET-seq data result from RNAPII pausing, and indicate that pausing followed by backtracking—which previously had been observed at promoter-proximal pauses[35]—is prevalent throughout the body of transcripts. Additionally, our studies indicate that Dst1-stimulated RNA cleavage has a strong sequence bias and that a slow step follows cleavage before transcription resumes.

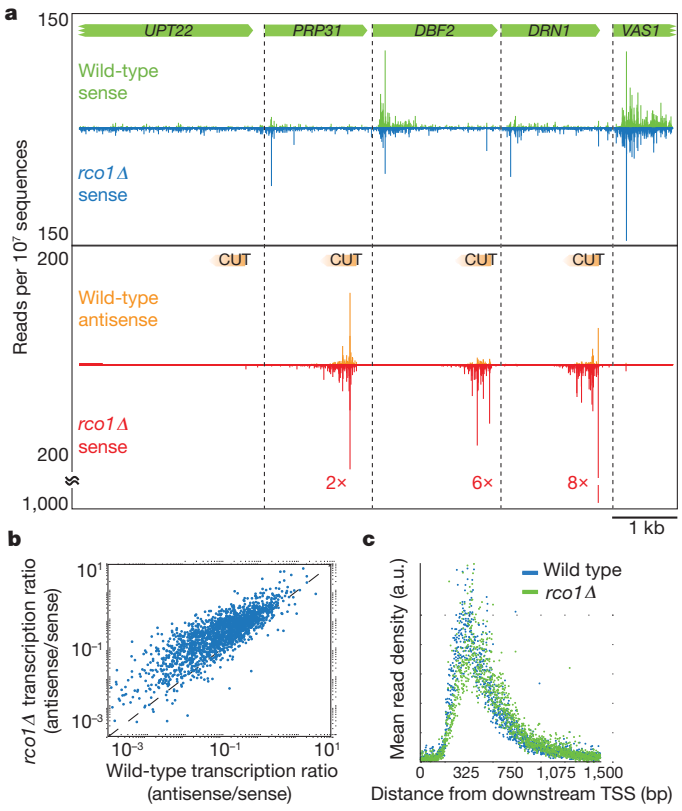


Figure 3 | Rco1 suppresses antisense transcription at divergent promoters. **a**, Examples of cryptic unstable transcripts (CUTs, orange data) upstream and antisense of *DBF2*, *DRN1* and *VAS1* promoters. The fold increase of CUT transcription in the *rco1Δ* strain is marked. **b**, The transcription ratio (antisense/sense) in the *rco1Δ* strain is plotted against the transcription ratio in the wild-type strain for each gene. **c**, A metagene analysis of well-expressed antisense transcription ($n = 171$, >1 read per bp).

RNAPII pause density peaks before the nucleosome dyad

The pauses observed in the *dst1Δ* strain reveal positions where RNAPII began to backtrack and, therefore, represent the primary point of transcriptional blockage. By analysing these pause positions, we can evaluate what induced RNAPII to backtrack. *In vitro*, nucleosomes induce RNAPII backtracking and TFIIS aids the progression of RNAPII through them[10,12]. *In vivo*, it is unknown whether nucleosomes interfere with transcription, as chromatin remodelling factors could greatly diminish the nucleosome barrier or remove nucleosomes before RNAPII arrival[37,38]. Global high-resolution measurements of steady-state nucleosome occupancy revealed that the first few nucleosomes after the transcription start site are phased and well positioned[23,39]. Thus, by correlating the relative density of RNAPII pauses with nucleosome positions, we can evaluate whether nucleosomes promote RNAPII pausing *in vivo*.

We compared the pause positions in the *dst1Δ* strain to the centre positions of nucleosomes using previously published data[23]. Notably, we saw marked peaks of mean pause density at each of the first four nucleosomes (Fig. 6). The precise position of the point of maximal RNAPII pausing at the +1 nucleosome is obscured because it is located just after the transcription start site where many nascent transcripts are too short for unique alignment to the genome. For the +2, +3 and +4 nucleosomes, however, the pause density peaks just before the nucleosome dyad axis (Fig. 6). As would be expected from RNAPII backtracking, the excess pause density before the nucleosome dyad in the wild-type strain is spread out over the upstream region (Supplementary Fig. 9).

Our finding that the peak in pause density occurs just before the nucleosome dyad is particularly remarkable as it is in excellent agreement with earlier biophysical measurements. Specifically, optical trapping studies that physically unwrapped the DNA of a nucleosome off the histone core observed that the dyad is the point where the strongest DNA–histone contacts are found[40]. Moreover, high-resolution optical trapping experiments that followed RNAPII transcribing through a nucleosome found that the RNAPII pause density peaked before the nucleosome dyad[10]. Taken together, the above observations provide strong evidence that nucleosomes do indeed present a barrier to elongating polymerases *in vivo* and that this barrier leads to polymerase pausing and backtracking.

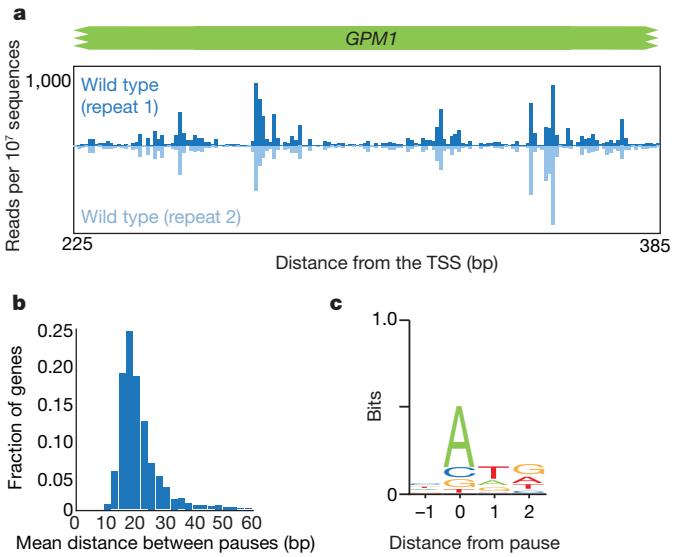


Figure 4 | Frequent RNAPII pausing throughout gene bodies. **a**, NET-seq data at the *GPM1* gene for biological replicates. **b**, A histogram of the mean distance between pauses for each well-expressed gene ($n = 1{,}006$, >2 reads per bp). **c**, The consensus sequence of the DNA coding strand surrounding pause sites found from all genes.

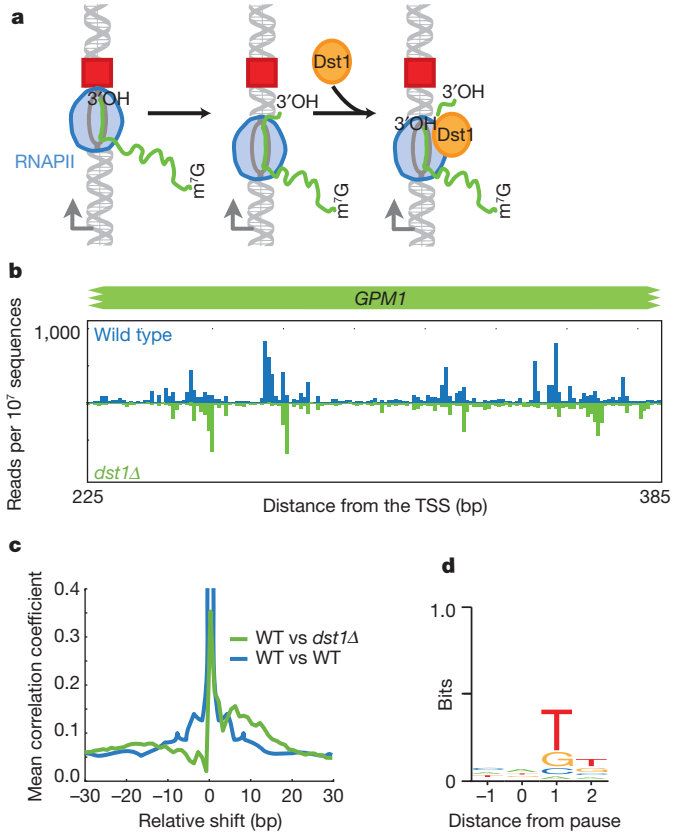


Figure 5 | Dst1 relieves RNAPII pausing after backtracking. **a**, A schematic describing an existing model for how RNAPII pauses at an obstacle (red square), backtracks and is induced to cleave its transcript through binding to Dst1 (refs 32, 33). **b**, A comparison of NET-seq data for wild-type and *dst1*Δ strains at the *GPM1* gene. **c**, Mean cross-correlation between the *dst1*Δ and wild-type data of well transcribed genes ($n = 770$, >2 reads per bp) (green line) was calculated by determining the Pearson's correlation coefficient at each gene between fixed *dst1*Δ data and shifted wild-type data followed by averaging over all genes. This analysis is compared to the mean autocorrelation of the wild-type data for well transcribed genes (blue line). **d**, The consensus sequence for all pauses observed in the *dst1*Δ strain.

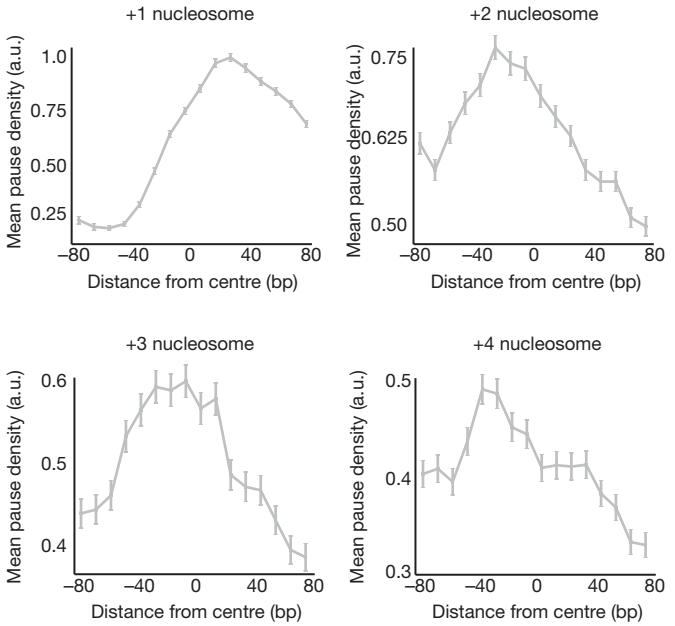


Figure 6 | Nucleosomes are a major barrier to transcription. Plot of mean pause densities in *dst1*Δ data relative to the first four nucleosomes after the transcription start site using available nucleosome positioning data[23]. Error bars represent one standard deviation.

Perspective

One of the major surprises in the transcription field in recent years has been the widespread observation of divergent transcription, revealing that the majority of promoters engage in canonical transcription in the sense direction along with the production of unstable transcripts in the antisense direction[2–4,15,21]. NET-seq provides an ideal tool to look at this phenomenon and uncovers several fundamental properties of divergent transcription. First, most promoters show a strong directionality favouring the sense transcript. Second, suppression of antisense transcripts is enforced by two distinct mechanisms: Rpd3S-mediated deacetylation that prevents antisense initiation, and an independent mechanism, previously characterized to involve the Nrd1–Nab3–Sen1 complex[41], that terminates antisense transcripts and shuttles them to the exosome for degradation. Interestingly, sense transcription may also use this termination mechanism, as our data showed an enrichment for transcripts at the 5′ end of genes that mirrors what we observed for antisense transcripts and complements observations that Nrd1 localizes to the 5′ end of genes[42]. Third, our observations indicate independence between the initiation of the sense and antisense transcripts. Specifically, we found only modest correlation between sense and antisense transcription levels. Moreover, even among the set of antisense transcripts that increased when *RCO1* is deleted, no increase in sense transcription levels was seen. These findings argue against models in which antisense transcription serves to promote sense transcription (for example, by unwinding DNA supercoils or by removing nucleosomes).

The potential for RNAP to pause has been apparent for decades, motivating interest in the mechanisms and regulatory roles of pausing in the process of transcription[7,9,11,12,15]. NET-seq provides the first in-depth view of pausing in a eukaryotic cell, revealing that transcription is punctuated by pauses throughout the body of all RNA messages. Taking into account both the abundance and magnitude of the pauses, we conclude that RNAPII spends comparable time in a paused state and moving forwards (Supplementary Fig. 10). We establish that nucleosomes induce pausing *in vivo*, and may be the major source of pausing considering that the increase in pause density at nucleosomes is comparable to the increase in nucleosome occupancy[23]. Our observation that pausing peaks at the nucleosome dyad reveal a striking similarity between our measurements and optical trap measurements, indicating that the physical forces observed in purified *in vitro* systems are at play in the cell. NET-seq's ability to follow the physical basis of transcription *in vivo*, allowing direct comparison with high-resolution *in vitro* measurements, may prove to be the most transformative aspect of this approach.

METHODS SUMMARY

Nascent RNA purification. All experiments were conducted using derivatives of yeast strain BY4741. Epitope-tagged Rpb3 (C-terminal 3×-Flag) was expressed from its endogenous locus. Deletion strains were made by standard PCR-based methods. Litres of log phase culture in YEPD were harvested by filtration and flash frozen by plunging into liquid nitrogen. Frozen cells were lysed cryogenically via six cycles of pulverization using a mixer mill.

Clarified and DNase-I-digested lysate was added to washed anti-Flag M2 affinity gel (Sigma Aldrich), incubated at 4 °C and nutated for 2.5 h. After washing, bound proteins were eluted twice with 2 mg ml^{-1} 3×-Flag peptide (Sigma Aldrich). RNA from the eluates was purified using the miRNeasy kit (Qiagen).

RNA linker ligation, cDNA synthesis and PCR. An RNA linker that was 5′ adenylated and 3′-end blocked with a dideoxy-C base (5′-CTGTAGGCACC ATCAAT, Integrated DNA Technologies) was ligated onto the 3′ end of the immunoprecipitated RNA based on a previously described strategy[43]. Ligation conditions (see Methods) were systematically optimized to maximize ligation efficiency to ~90% to ensure that the majority of the input RNA was ligated and thus avoiding any bottleneck biases.

cDNA synthesis and sequencing was performed as described with a few modifications[19]. The sequencing primer binding site was positioned so that sequencing would start at the 3′ end.

Comparing pause densities to nucleosome positions. Nucleosome positions[23] were assigned as +1, +2, +3 etc according to their position relative to transcription start sites. The mean pause density (MPD) relative to a particular nucleosome was determined by the number of pauses observed at that position (N_p) divided by the total number of opportunities it could be observed there (N_o):

$$\mathrm{MPD}_k(x) = \left(\frac{N_p}{N_o}\right)_x = \frac{\sum_{i}^{\text{all genes}} g_i(y)}{\sum_{i}^{\text{genes with TSS} < y} 1}$$

$$y = n_i^k + x$$

where k is the nucleosome number, $g(y)$ is the binary function indicating whether a pause occurs at y, and n_i^k are the centre nucleosome positions. The error of the pause density was calculated via the standard deviation of the binomial distribution:

$$\frac{\sqrt{N_p\left(1 - N_p/N_o\right)}}{N_o}$$

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 1 June; accepted 8 November 2010.

1. Moore, M. J. & Proudfoot, N. J. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* **136,** 688–700 (2009).
2. Preker, P. *et al.* RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322,** 1851–1854 (2008).
3. Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457,** 1033–1037 (2009).
4. Neil, H. *et al.* Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* **457,** 1038–1042 (2009).
5. Rougvie, A. E. & Lis, J. T. The RNA polymerase II molecule at the 5′ end of the uninduced *hsp70* gene of *D. melanogaster* is transcriptionally engaged. *Cell* **54,** 795–804 (1988).
6. Proshkin, S., Rahmouni, A. R., Mironov, A. & Nudler, E. Cooperation between translating ribosomes and RNA polymerase in transcription elongation. *Science* **328,** 504–508 (2010).
7. Kassavetis, G. A. & Chamberlin, M. J. Pausing and termination of transcription within the early region of bacteriophage T7 DNA *in vitro*. *J. Biol. Chem.* **256,** 2777–2786 (1981).
8. Shaevitz, J. W., Abbondanzieri, E. A., Landick, R. & Block, S. M. Backtracking by single RNA polymerase molecules observed at near-base-pair resolution. *Nature* **426,** 684–687 (2003).
9. Herbert, K. M. *et al.* Sequence-resolved detection of pausing by single RNA polymerase molecules. *Cell* **125,** 1083–1094 (2006).
10. Hodges, C., Bintu, L., Lubkowska, L., Kashlev, M. & Bustamante, C. Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. *Science* **325,** 626–628 (2009).
11. Kireeva, M. L. & Kashlev, M. Mechanism of sequence-specific pausing of bacterial RNA polymerase. *Proc. Natl Acad. Sci. USA* **106,** 8900–8905 (2009).
12. Kireeva, M. L. *et al.* Nature of the nucleosomal barrier to RNA polymerase II. *Mol. Cell* **18,** 97–108 (2005).
13. Kim, T. H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436,** 876–880 (2005).
14. Lefrançois, P. *et al.* Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing. *BMC Genomics* **10,** 37 (2009).
15. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322,** 1845–1848 (2008).
16. Rodríguez-Gil, A. *et al.* The distribution of active RNA polymerase II along the transcribed region is gene-specific and controlled by elongation factors. *Nucl. Acids Res.* **38,** 4651–4664 (2010).
17. Cai, H. & Luse, D. S. Transcription initiation by RNA polymerase II *in vitro*. Properties of preinitiation, initiation, and elongation complexes. *J. Biol. Chem.* **262,** 298–304 (1987).
18. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
19. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–223 (2009).
20. Markham, R. & Smith, J. D. The structure of ribonucleic acids. I. Cyclic nucleotides produced by ribonuclease and by alkaline hydrolysis. *Biochem. J.* **52,** 552–557 (1952).
21. Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322,** 1849–1851 (2008).
22. Seila, A. C., Core, L. J., Lis, J. T. & Sharp, P. A. Divergent transcription: a new feature of active promoters. *Cell Cycle* **8,** 2557–2564 (2009).
23. Weiner, A., Hughes, A., Yassour, M., Rando, O. J. & Friedman, N. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Res.* **20,** 90–100 (2010).
24. Pokholok, D. K. *et al.* Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122,** 517–527 (2005).
25. Carrozza, M. J. *et al.* Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* **123,** 581–592 (2005).
26. Keogh, M. C. *et al.* Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. *Cell* **123,** 593–605 (2005).
27. Li, B. *et al.* Histone H3 lysine 36 dimethylation (H3K36me2) is sufficient to recruit the Rpd3s histone deacetylase complex and to repress spurious transcription. *J. Biol. Chem.* **284,** 7970–7976 (2009).
28. Pinskaya, M., Gourvennec, S. & Morillon, A. H3 lysine 4 di- and tri-methylation deposited by cryptic transcription attenuates promoter activation. *EMBO J.* **28,** 1697–1707 (2009).
29. Govind, C. K. *et al.* Phosphorylated Pol II CTD recruits multiple HDACs, including Rpd3C(S), for methylation-dependent deacetylation of ORF nucleosomes. *Mol. Cell* **39,** 234–246 (2010).
30. Krogan, N. J. *et al.* Methylation of histone H3 by Set2 in *Saccharomyces cerevisiae* is linked to transcriptional elongation by RNA polymerase II. *Mol. Cell. Biol.* **23,** 4207–4218 (2003).
31. Nudler, E., Mustaev, A., Lukhtanov, E. & Goldfarb, A. The RNA-DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase. *Cell* **89,** 33–41 (1997).
32. Izban, M. G. & Luse, D. S. Factor-stimulated RNA polymerase II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates. *J. Biol. Chem.* **267,** 13647–13655 (1992).
33. Reines, D., Conaway, R. C. & Conaway, J. W. Mechanism and regulation of transcriptional elongation by RNA polymerase II. *Curr. Opin. Cell Biol.* **11,** 342–346 (1999).
34. Kulish, D. & Struhl, K. TFIIS enhances transcriptional elongation through an artificial arrest site *in vivo*. *Mol. Cell. Biol.* **21,** 4162–4168 (2001).
35. Nechaev, S. *et al.* Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. *Science* **327,** 335–338 (2010).
36. Sigurdsson, S., Dirac-Svejstrup, A. B. & Svejstrup, J. Q. Evidence that transcript cleavage is essential for RNA polymerase II transcription and cell viability. *Mol. Cell* **38,** 202–210 (2010).
37. Li, B., Carey, M. & Workman, J. The role of chromatin during transcription. *Cell* **128,** 707–719 (2007).
38. Petesch, S. J. & Lis, J. T. Rapid, transcription-independent loss of nucleosomes over a large chromatin domain at Hsp70 loci. *Cell* **134,** 74–84 (2008).
39. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458,** 362–366 (2009).
40. Hall, M. A. *et al.* High-resolution dynamic mapping of histone-DNA interactions in a nucleosome. *Nature Struct. Mol. Biol.* **16,** 124–129 (2009).
41. Arigo, J. T., Eyler, D. E., Carroll, K. L. & Corden, J. L. Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. *Mol. Cell* **23,** 841–851 (2006).
42. Vasiljeva, L., Kim, M., Mutschler, H., Buratowski, S. & Meinhart, A. The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. *Nature Struct. Mol. Biol.* **15,** 795–804 (2008).
43. Unrau, P. J. & Bartel, D. P. RNA-catalysed nucleotide synthesis. *Nature* **395,** 260–263 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Guthrie, N. Krogan, S. Luo, G. Schroth, J. Steitz and K. Yamamoto for advice and discussions; D. Breslow, P. Fordyce, A. Frost, J. Huff, M. Kampmann and M. Pufall for critical comments on the manuscript; C. Chu and N. Ingolia for help with sequencing and analysis; and S. Rouskin for help developing the ligation protocol. This research was supported by the Damon Runyon Cancer Research Foundation (DRG-1997-08 to L.S.C.) and by the Howard Hughes Medical Institute (to J.S.W.)

Author Contributions L.S.C. and J.S.W. designed the experiments; L.S.C. performed the experiments and analysed the data; and L.S.C. and J.S.W. interpreted the results and wrote the manuscript.

Author Information Raw sequencing data and processed data are available for download at http://www.ncbi.nlm.nih.gov/geo/ via GEO accession number GSE25107. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.S.W. (weissman@cmp.ucsf.edu).

METHODS

Strain construction. All experiments were conducted using derivatives of yeast strain BY4741. Epitope-tagged Rpb3 (C-terminal 3×-Flag) was expressed from its endogenous locus. Deletion strains were made by standard PCR-based methods.

Extract and total RNA preparation. Yeast strains were grown in YEPD at 30 °C with shaking from an initial optical density (OD) of 0.1 to mid-log phase with an OD of 0.6–0.8. Two litres of yeast culture were harvested in turn by filtration onto 0.45-μm-pore-size nitrocellulose filters (Whatman). The culture was scrapped off the filter with a spatula pre-chilled by liquid nitrogen and flash frozen by plunging into liquid nitrogen. Frozen cells were pulverized for six cycles, each of 3 min at 15 Hz, on a Retsch MM301 mixer mill. Sample chambers were pre-chilled in liquid nitrogen and re-chilled between each pulverization cycle.

One gram of ground cells (~ 1 l at 0.7 OD) was added to 5 ml of ice-cold lysis buffer (20 mM HEPES, pH 7.4, 110 mM KOAc, 0.5% Triton X-100, 0.1% Tween 20, 10 mM $MnCl_2$, 50 U ml^{-1} SUPERase•In (Ambion)) supplemented with protease inhibitor cocktail (1× Complete, EDTA-free, Roche). The experiment using α-amanitin included 10 μg ml^{-1} α-amanitin (Sigma Aldrich) in the lysis buffer. After re-suspending the lysate by pipetting, 660 units of DNase I (Promega, RQ1 RNase-Free DNase) was added and incubated for 20 min on ice. The lysate was then clarified by centrifugation at 4 °C at 20,000*g* for 10 min. The supernatant is reserved for immunoprecipitation.

Two-hundred microlitres of clarified lysate is reserved for total RNA purification which was done by the hot acid phenol method. Typical yields were 20 μg.

Native affinity purifications of RNAPII. 0.5 ml of Anti-Flag M2 Affinity Gel (Sigma Aldrich) was washed twice with lysis buffer. The clarified lysate was added to the washed gel, incubated at 4 °C and nutated for 2.5 h. The immunoprecipitation was washed 4 × 10 ml with wash buffer (20 mM HEPES, pH 7.4, 110 mM KOAc, 0.5% Triton X-100, 0.1% Tween 20, 50 U ml^{-1} SUPERase•In (Ambion), 1 mM EDTA). Bound proteins were eluted twice with 150 μl elution buffer (20 mM HEPES, pH 7.4, 110 mM KOAc, 0.5% Triton X-100, 0.1% Tween 20) with 2 mg ml^{-1} 3×-Flag peptide (Sigma Aldrich). RNA from the combined eluates was purified using the miRNeasy kit (Qiagen, 217004). A typical yield from approximately one litre of log-phase yeast culture was 3 μg.

mRNA purification and fragmentation. Polyadenylated mRNA was purified from 50 μg total RNA using magnetic oligo-dT DynaBeads (Invitrogen). Purified RNA was eluted in 20 μl 10 mM Tris, pH 7.0. The purified mRNA was mixed with an equal volume of 2× alkaline fragmentation solution (2 mM EDTA, 10 mM Na_2CO_3, 90 mM $NaHCO_3$, pH ≈ 9.3) and incubated for 5 min at 95 °C. These conditions yielded lightly fragmented RNA of size distribution similar to that of the nascent RNA. The fragmentation reaction was stopped by the addition 0.56 ml of ice-cold precipitation solution (final 300 mM NaOAc pH 5.5, plus GlycoBlue (Ambion) as a co-precipitant) and RNA was purified by a standard isopropanol precipitation, as follows: after adding 650 μl of isopropanol, samples were placed at −30 °C for at least 30 min. Precipitated RNA was pelleted by centrifugation at 4 °C at 20,000*g* for 30 min. The pellet was air dried after a quick wash with 80% ethanol and then re-suspended in 10 mM Tris pH 7.0.

A total of 6.4 μg of fragmented mRNA was dephosphorylated in a 50 μl reaction with 1× T4 polynucleotide kinase buffer without ATP, 0.5 U SUPERase•In (Ambion) and 22.5 units T4 polynucleotide kinase (NEB). The dephosphorylation reaction was incubated at 37 °C for 1 h followed by 10 min at 75 °C for enzyme heat inactivation. RNA was precipitated with GlycoBlue by standard methods (see above).

RNA linker ligation, fragmentation and size selection. An RNA linker that was 5′ adenylated and 3′-end blocked with a dideoxy-C base (5′-CTGTAGGCACCA TCAAT, Integrated DNA Technologies) was ligated onto the 3′ end of the immunoprecipitated RNA, the fragmented mRNA and a synthetic 28-base RNA oligonucleotide (oNTI199, 5′-AUGUACACGGAGUCGACCCGCAACG CGA) similarly to what has been described[43]. Specifically, 3 μg of each RNA sample was broken into three reactions and diluted to 10 μl with 10 mM Tris, pH 7.0. After a brief denaturation the reactions were brought to 20 μl with a buffer that gave final concentrations of 12% PEG8000, 50 ng μl^{-1} linker, 1× T4 Rnl2, truncated reaction buffer and 2 units μl^{-1} of T4 Rnl2, truncated (NEB). The reaction was incubated at 37 °C for 3 h. Ligation conditions were systematically optimized to maximize ligation efficiency to ~90% to ensure that the majority of the input RNA was ligated.

Fragmentation of the ligated samples allowed for the final DNA library to contain inserts of a narrow range to reduce any length biases of downstream enzymatic reactions. EDTA was added to all reactions for a final concentration of 17 mM. 20 μl of 2× alkaline fragmentation solution (2 mM EDTA, 10 mM Na_2CO_3, 90 mM $NaHCO_3$, pH ≈ 9.3) was added to each reaction and incubated at 95 °C for 30 min. The reactions were stopped by the addition of 0.56 ml of ice-cold precipitation solution (final 300 mM NaOAc pH 5.5, plus GlycoBlue (Ambion) as a co-precipitant), followed by a standard isopropanol precipitation (see above).

The ligated and fragmented samples were size-selected by gel electrophoresis. The purified reactions along with the oNTI199 RNA oligonucleotide was mixed with 2× Novex TBE-Urea sample prep buffer (Invitrogen) and briefly denatured, then loaded on a Novex denaturing 15% polyacrylamide TBE-urea gel (Invitrogen) and run according to the manufacturer's instructions. The gel was stained with SYBR Gold (Invitrogen) and the 35–85-nucleotide region was excised. The gel was physically disrupted and either allowed to soak overnight in gel elution buffer (300 mM NaOAc pH 5.5, 1 mM EDTA, 0.1 U μl^{-1} SUPERase·In) or incubated in 200 μl of water treated with diethylpyrocarbonate (DEPC) for 10 min at 70 °C. The gel debris was removed from the water or buffer using a Spin-X column (Corning) and RNA was precipitated with GlycoBlue as a co-precipitant using standard methods.

cDNA synthesis. cDNA synthesis was performed as described with a few modifications[19]. The primer used for reverse transcription was oLSC003 (5′-pTCG TATGCCGTCTTCTGCTTG•-AATGATACGGCGACCACCGATCCGACGAT CATTGATGGTGCCTACAG) where the initial 'p' indicates 5′ phosphorylation and '•' indicates the following spacer added for increased flexibility, 18 carbon spacer molecule-CACTCA-18 carbon spacer molecule. Efficient circularization of the RT product was performed as described[19] with CircLigase (Epicentre) according to the manufacturer's directions. Any ligation bias at this step is averaged out as the random fragmentation leaves a range of 5′ ends for each 3′ end. The PCR was performed directly on the circularized product as described[19], resulting in DNA with Illumina cluster generation sequences on each end and a sequencing primer binding site positioned so that sequencing would start at the 3′ end. DNA was purified from a PCR reaction that had not reached saturation and was quantified using the Agilent BioAnalyser High Sensitivity DNA assay. DNA was then sequenced on the Illumina Genome Analyser 2 according to the manufacturer's instructions, using 4–6 pM template for cluster generation and sequencing primer oLSC006 (5′-TCCGACGATCATTGATGGTGCCTACAG).

Data analysis. Data analysis was performed using scripts written in Python 2.6 that are available upon request.

Sequencing analysis. Image data obtained by the Illumina Genome Analyser 2 was analysed using the GAPipeline to extract raw sequences. Matrix and phasing parameters were estimated from a ϕX control lane.

Sequence alignment. Raw sequences 40 bases long were composed of the cDNA of the fragmented RNA sequence. For RNA fragments smaller than 40 bases, the sequence is followed by part of the 5′ Illumina linker sequence which was removed *in silico*. Alignments to the yeast genome were performed by the alignment program, Bowtie 0.12.0[44] (http://bowtie-bio.sourceforge.net/). Bowtie settings were chosen so that three mismatches were allowed and alignments were required to be unique. The shortest sequenced fragments were approximately 18 nucleotides due to the RNA size selection step after ligation and random fragmentation. Eighteen-base-pair sequences would occur by chance every 6.9×10^{10} bp, which is sufficiently rare for 18-bp sequences to be generally uniquely aligned to the 1.2×10^{7} bp yeast genome. Alignments were first performed against tRNA and rRNA sequences to remove them. The remaining sequences were aligned against a recent version of the yeast genome downloaded from the Saccharomyces Genome Database (SGD, http://www.yeastgenome.org/) on 11 October 2009. Statistics on sequence alignments are reported in Supplementary Table 1.

Quantifying antisense and sense transcription levels. At tandem promoters sense transcription was determined using available annotated transcription start sites[3]. To allow for the error involved in these transcription start site measurements, we calculated the sum of the read density in 500-nucleotide windows for the first 700 bases after the transcription start site and chose the highest sum. The antisense transcription was determined by starting 100 bases upstream of the transcription start site and the read density sum in 500-nucleotide-wide windows was calculated for the subsequent 1,000 bases. The highest sum was used for downstream analysis.

Metagene analysis. Each gene included in the analysis is normalized by the mean number of reads in a 400-bp window beginning 100 bases downstream from the transcription start site. A mean read density (MRD) is then calculated for each position over all genes as described below.

$$\mathrm{MRD}(i) = \frac{\sum_{j}^{\text{all genes}} \left(\frac{r_i^j}{\sum_{i=100}^{500} r_i^j \Big/ 400} \right)}{\sum_{j}^{\text{all genes}} 1}$$

where r_i^j are the reads for the jth gene at the ith position after the transcription start site.

Extracting pause positions. Pauses were identified in previously annotated transcription units[3] of well-expressed genes. Pauses were defined as having reads

higher than three standard deviations above the mean of the surrounding 200 nucleotides which do not contain pauses. Pauses were required to have at least four reads regardless of the gene's sequencing coverage. Sequence consensus was calculated by WebLogo 3 (http://weblogo.threeplusone.com/)[45].

Comparing pause densities to nucleosome positions. Nucleosome positions[23] were assigned as +1, +2, +3 etc according to their position relative to transcription start sites. The mean pause density (MPD) relative to a particular nucleosome was determined by the number of pauses observed at that position (N_p) divided by the total number of opportunities it could be observed there (N_o):

$$\mathrm{MPD}_k(x) = \left(\frac{N_p}{N_o}\right)_x = \frac{\sum_i^{\text{all genes}} g_i(y)}{\sum_i^{\text{genes with TSS} < y} 1}$$

$$y = n_i^k + x$$

where k is the nucleosome number, $g(y)$ is the binary function indicating whether a pause occurs at y, and n_i^k are the centre nucleosome positions. For the +2 and +3 nucleosomes, the number of pause opportunities was uniform at every position and was simply the number of genes included in the analysis. The +1 nucleosome analysis required that the number of pause opportunities at each position represent the number of genes where that position occurs after the transcription start site. The error of the pause density was calculated via the standard deviation of the binomial distribution

$$\frac{\sqrt{N_p\left(1 - N_p/N_o\right)}}{N_o}$$

The densities were then binned by averaging across windows ten nucleotides wide. The error for each bin was calculated by computing the sum of the variances of the binned measurements and calculating the square root.

44. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
45. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–1190 (2004).

LETTER

doi:10.1038/nature09694

Supermassive black holes do not correlate with galaxy disks or pseudobulges

John Kormendy[1], R. Bender[2,3] & M. E. Cornell[1]

The masses of supermassive black holes are known to correlate with the properties of the bulge components of their host galaxies[1–5]. In contrast, they seem not to correlate with galaxy disks[1]. Disk-grown 'pseudobulges' are intermediate in properties between bulges and disks[6]; it has been unclear whether they do[1,5] or do not[7–9] correlate with black holes in the same way that bulges do. At stake in this issue are conclusions about which parts of galaxies coevolve with black holes[10], possibly by being regulated by energy feedback from black holes[11]. Here we report pseudobulge classifications for galaxies with dynamically detected black holes and combine them with recent measurements of velocity dispersions in the biggest bulgeless galaxies[12]. These data confirm that black holes do not correlate with disks and show that they correlate little or not at all with pseudobulges. We suggest that there are two different modes of black-hole feeding. Black holes in bulges grow rapidly to high masses when mergers drive gas infall that feeds quasar-like events. In contrast, small black holes in bulgeless galaxies and in galaxies with pseudobulges grow as low-level Seyfert galaxies. Growth of the former is driven by global processes, so the biggest black holes coevolve with bulges, but growth of the latter is driven locally and stochastically, and they do not coevolve with disks and pseudobulges.

In Fig. 1b, we show the well-known correlation[1,5] between dynamically measured black-hole masses, $M_\bullet$, and the absolute magnitudes of elliptical galaxies (Fig. 1b, black points) and the bulge parts of disk galaxies (Fig. 1b, red points). The correlation has $\chi^2 = 12$ per degree of freedom, implying moderate intrinsic scatter. This result and a tighter correlation[2–5] between $M_\bullet$ and host velocity dispersion, σ (Fig. 2), motivate the idea that black holes and host bulges evolve together and regulate each other[10,11]. All new results in this paper stand in contrast to these two correlations.

In Fig. 1a, we plot $M_\bullet$ versus the K-band absolute magnitude, M_K, of only the disk part of the host galaxy, with the bulge luminosity removed. We conclude that black holes do not correlate with galaxy disks. This confirms an earlier conclusion[1] based on the more indirect observation that black holes do not correlate with total (bulge plus disk) luminosities of disk galaxies. In Fig. 1a, the colour of each point—which indicates bulge type (see below)—is irrelevant. A least-squares fit to the red and blue points has correlation coefficient $r = 0.41$. Errors at 1 s.d. imply $\chi^2 = 81$ per degree of freedom: the data do not respect the weak anticorrelation. The green points for pure-disk (that is, completely bulgeless) galaxies further confirm the large scatter and lack of correlation.

In Fig. 1c, we plot $M_\bullet$ versus M_K for pseudobulge components with disk light removed. They are also included in Fig. 1b (light blue) to show how they compare with classical bulges and elliptical galaxies. Pseudobulges required explanation, as follows.

Much work over several decades has shown that the high-stellar-density central components in disk galaxies—all of which used to be called 'bulges'—come in two varieties. How to distinguish them is discussed in Supplementary Information. The difference was first found observationally but is now understood to be a result of fundamentally different formation mechanisms[6].

Classical bulges (red points in Figs 1 and 2) are indistinguishable from ellipticals in their structure, velocity distributions and parameters. Our well-developed picture is that they formed by galaxy mergers[13] in our hierarchically clustering universe. Mergers are discrete events that are separated by long 'dead times'. They occur on short timescales, approximately equal to the crossing time of the merging galaxies. Gravitational torques scramble disks into ellipticals[13] and dump large quantities of gas into the centre. Observations[14] and theory[15] suggest that the process feeds both starbursts and black holes and causes the latter to grow rapidly in quasar-like events.

Pseudobulges[6] (blue points in Figs 1 and 2) are observed to be more disk-like than classical bulges. They are believed to form more gently by the gradual internal redistribution of angular momentum in quiescent galaxy disks. The driving agents are non-axisymmetries such as bars. One result is the gradual build-up of a high-density central component that can be recognized (Supplementary Information) because it remains disk-like. These components are called 'pseudobulges' to emphasize their different formation mechanism without forgetting that they look superficially like—and are commonly confused with—classical bulges. The difference from bulges that is most relevant here is that the gradual gas infall that builds pseudobulges may concurrently provide less black-hole feeding and thus may drive slower black-hole growth. One purpose of this paper is to contrast black hole/bulge and black hole/pseudobulge correlations to look for clues about black-hole growth mechanisms and the consequent coevolution (or not) of black holes with host galaxies.

With this context, we can interpret Fig. 1. Galaxies that contain classical bulges are consistent with the correlations for elliptical galaxies except for one discrepant object (the bulge-dominated S0 galaxy NGC 4342). The implication is that classical bulges and ellipticals coevolve with black holes in the same way. For that coevolution, it is irrelevant that bulges are now surrounded by disks whereas ellipticals are not.

We reach a different conclusion for pseudobulges on the basis of new classifications and measurements of pseudobulge-to-total luminosity ratios for all disk galaxies in our sample[5] that have dynamical black-hole detections (J.K., manuscript in preparation; see Supplementary Information for a list). A conservative interpretation of Fig. 1b is that pseudobulges are roughly consistent with the correlation for classical bulges and ellipticals but have much more scatter. In particular, some deviate from the correlation for ellipticals in having smaller black holes. This was not seen in some previous work[1,5] because samples were small and because many pseudobulges in black-hole galaxies had not been classified. But, as published samples have grown larger, the hints have grown stronger that pseudobulges do not correlate with black holes in the same way as classical bulges[7–9]. We confirm these hints. Particularly compelling is the fact that our galaxies and a new sample of black-hole detections based on water masers[9] have no overlap and independently lead to the same conclusion.

Figure 1c shows the pseudobulges without guidance from the red and black points. The sample is small, but we have enough dynamic

[1]Department of Astronomy, University of Texas at Austin, 1 University Station, Austin, Texas 78712-0259, USA. [2]Max-Planck-Institut für Extraterrestrische Physik, Giessenbachstrasse, D-85748 Garching-bei-München, Germany. [3]Universitäts-Sternwarte, Scheinerstrasse 1, D-81679 München, Germany.

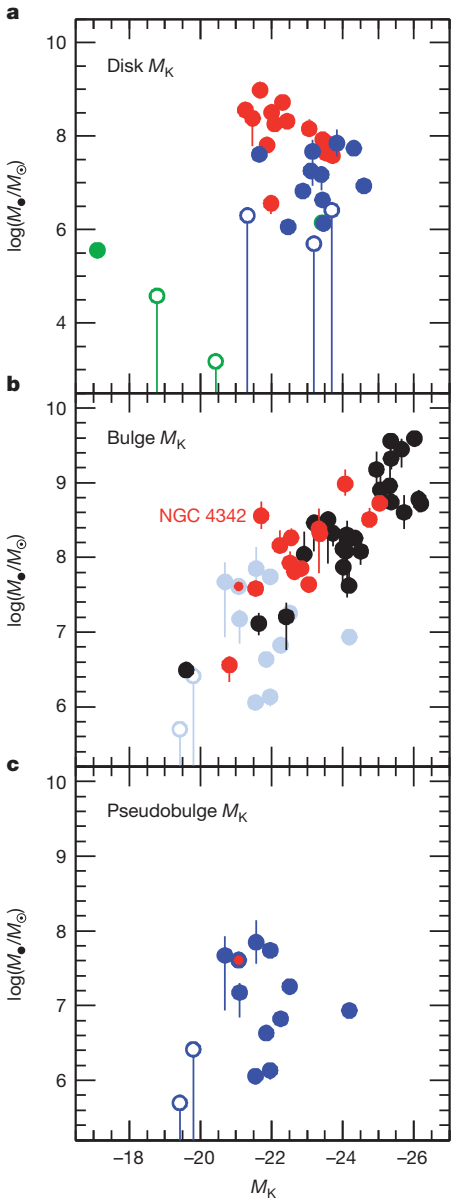


Figure 1 | Correlations of dynamically measured black-hole masses with the luminosities of different parts of their host galaxies. Here M_K is the K-band (2.2-μm) absolute magnitude of the disk component with bulge light removed (**a**), of the bulge with disk light removed (**b**) and of the pseudobulge with disk light removed (**c**). All plotted data are published elsewhere; parameters and sources are discussed in Supplementary Information, and those for disk galaxies are tabulated there. Elliptical galaxies are plotted in black, classical bulges are plotted in red and pseudobulges are plotted in dark blue. One galaxy with a dominant pseudobulge but with a possible small classical bulge (NGC 2787) is plotted with a blue symbol that has a red centre. In least-squares fits, it is included with the pseudobulges. Error bars, 1 s.d. In **b**, the red and black points show a good correlation between $M_\bullet$ and bulge luminosity: a symmetric, least-squares fit[4] of a straight line has $\chi^2 = 12.1$ per degree of freedom and a Pearson correlation coefficient of $r = -0.82$. (All χ^2 values quoted in this paper are per degree of freedom.) In contrast, in **a** the red and blue points together confirm a previous result[1] that black holes do not correlate with disks: $\chi^2 = 81$ and $r = 0.41$. Green points are for galaxies that contain neither a classical bulge nor a pseudobulge but only a nuclear star cluster, that is, pure-disk galaxies. They are not included in the above fit, but they strengthen our conclusion. Similarly, in **c** the blue points for pseudobulges show no correlation: $\chi^2 = 63$ and $r = 0.27$. In all panels, galaxies that have only $M_\bullet$ limits are plotted with open symbols; they were chosen to increase our dynamic range. They too support our conclusions. This figure uses K-band magnitudes to minimize effects of star formation and internal absorption, but in Supplementary Information we show that Fig. 1 looks essentially the same for V-band (0.55-μm) magnitudes.

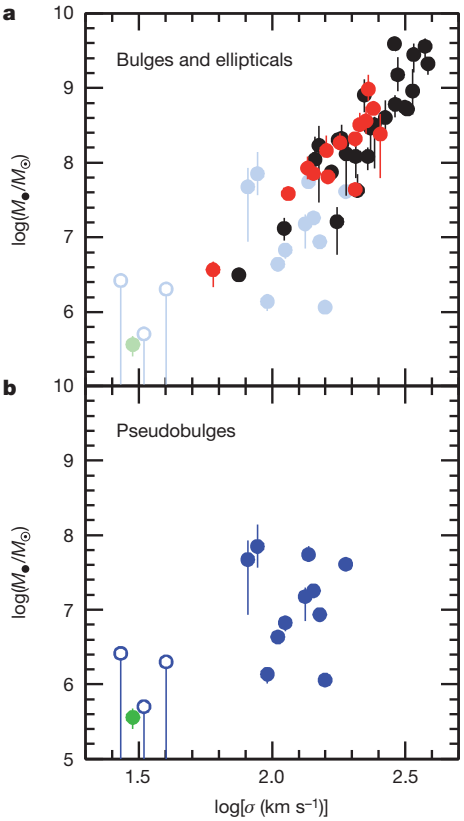


Figure 2 | Correlation of dynamically measured black-hole masses with the velocity dispersions of their host galaxies. Black points are for elliptical galaxies, red points are for classical bulges, blue points are for pseudobulges and the green point is for a nuclear star cluster. Data sources are given in Supplementary Information. Error bars, 1 s.d. The red and black points show the well known $M_\bullet$–σ correlation[2–5]: $\chi^2 = 5.0$ per degree of freedom and $r = 0.89$. Reducing χ^2 to 1.0 implies that the intrinsic scatter in $\log(M_\bullet/M_\odot)$ at fixed σ is 0.26, consistent with previous derivations[4,5]. This is the tightest correlation between black holes and host galaxy properties and the one that most motivates the idea that black holes and bulges coevolve. In contrast, the blue points for pseudobulges show no correlation: $\chi^2 = 10.4$ and $r = -0.08$. This extends suggestions[7–9] that the $M_\bullet$–σ relation for pseudobulges is different from that for classical bulges and elliptical galaxies.

range to conclude that we see no correlation at all. The cumulative amount of black-hole growth is not extremely different in classical bulges and pseudobulges, but there is no sign in the correlations that black-hole feeding has affected the pseudobulges.

The second and more compelling black hole/host galaxy correlation is the one between $M_\bullet$ and the velocity dispersion, σ, of the stars at radii where they do not feel the black hole gravitationally[2–5]. Here σ is averaged inside the 'effective radius', r_e, that contains half of the bulge light. Figure 2 shows this correlation.

As is well known, ellipticals and classical bulges share the same tight correlation. But as in Fig. 1, pseudobulges at best show a much larger scatter (Fig. 2a). Without the guidance of the red and black points (Fig. 2b), they show essentially no correlation. Larger samples that reach smaller values of $M_\bullet$ may show a weak relationship[16–20]. But we conclude that classical bulges and pseudobulges show very different correlations with $M_\bullet$. Those for classical bulges are tight enough to suggest coevolution. Whether pseudobulges correlate with $M_\bullet$ with large scatter or not at all, the weakness of any correlation ($r = -0.08$ here) makes no compelling case that pseudobulges and black holes coevolve, beyond the obvious expectation that it is easier for bigger black holes and bigger pseudobulges to grow in bigger galaxies that contain more fuel.

From the point of view of galaxy formation by hierarchical clustering, pseudobulge galaxies are already pure-disk galaxies[12]. Even more

extreme galaxies that contain neither a classical bulge nor a pseudo-bulge can, in some cases, also contain black holes. Some have active galactic nuclei (AGNs), although these are rare in bulgeless galaxies[21]. And black holes with $M_\bullet = 10^4 M_\odot$–$10^6 M_\odot$ ($M_\odot$, solar mass) have been discovered with confidence in bulgeless (and pseudobulgeless) galaxies[18,21]. The most extreme example is NGC 4395, an Sm galaxy that contains only a tiny, globular-cluster-like nucleus but that has an AGN powered by a black hole[22] with $M_\bullet = (3.6 \pm 1.1) \times 10^5 M_\odot$ well measured by reverberation mapping[23]. It is the green point with the small error bar in Fig. 2. Other such objects include the Sd galaxy NGC 3621[24] and the spheroidal galaxy POX 52, which contains an AGN powered by a black hole of mass $M_\bullet \approx 10^5 M_\odot$ (refs 18, 25). Also, it is likely that low-$M_\bullet$ AGN samples include bulgeless galaxies[16–19], although some are distant and not well resolved.

The lack of correlation of black holes with disks and pseudobulges plus the discovery of black holes in bulgeless galaxies together are critical clues to black-hole feeding mechanisms. They motivate the following hypothesis: we suggest that there are two fundamentally different feeding mechanisms for black holes.

The traditional mechanism is rapid feeding during major mergers, when large amounts of gas fall into galaxy centres. We follow suggestions[14,15] that, at some, perhaps late, stage in the merger, the black hole grows quickly in a quasar-like event. These are the growth episodes that dominate the waste mass budget of quasar activity[15,26] that is well explained by the observed density of the biggest black holes in the universe[27]. In this mode, black-hole and galaxy growth are controlled by the same global processes, and these, we suggest, result in black hole/bulge coevolution. If black holes and galaxy formation ever regulate each other[11], this is the likely setting in which it happens.

In contrast, the nuclear activity in bulgeless galaxies is, by and large, weak[21]. Big pseudobulges (NGC 1068 is the highest-luminosity one in Fig. 1) can host classical Seyfert nuclei, but there is no sign that these affect global galaxy structure. NGC 1068 is a prototypical oval galaxy[6] with a prominent pseudobulge that, we believe, formed slowly by inward gas transport in spite of episodic nuclear activity. We suggest that the second mode of black-hole growth is such weak nuclear activity driven stochastically by local processes that feed gas from $\sim 10^2$ pc, where it make a pseudobulge, in to the black hole. The processes are not understood in detail. But exactly this kind of feeding mode is proposed and modelled in ref. 26; the models suggest that this kind of feeding does not affect galaxy formation. Differences in black-hole growth and coevolution with host galaxies have also been proposed in studies of the lowest-$M_\bullet$ black holes detected mainly through AGN activity[16,18]. The same picture can also be reached by studying AGN demographics[28]. Other recent papers[29,30] further explore local processes of black-hole feeding. Figures 1 and 2 tell us that this mode involves little or no coevolution of black holes with any component of the host galaxy.

The seed black holes that grew into today's supermassive black holes are not securely identified. But the smallest black holes that are grown by means of local feeding plausibly remain most like those seeds. Because mergers of disk galaxies are believed to make bulges and ellipticals, we suggest that the small black holes that are grown by local processes are the seeds of the generally larger black holes that are grown in part by mergers.

Received 12 July; accepted 19 November 2010.

1. Kormendy, J. & Gebhardt, K. in *Proc. 20th Texas Symp. Relativ. Astrophys.* (eds Wheeler, J. C. & Martel, H.) 363–381 (American Institute of Physics, 2001).
2. Ferrarese, L. & Merritt, D. A fundamental relation between supermassive black holes and their host galaxies. *Astrophys. J.* **539,** L9–L12 (2000).
3. Gebhardt, K. *et al.* A relationship between nuclear black hole mass and galaxy velocity dispersion. *Astrophys. J.* **539,** L13–L16 (2000).
4. Tremaine, S. *et al.* The slope of the black hole mass versus velocity dispersion correlation. *Astrophys. J.* **574,** 740–753 (2002).
5. Gültekin, K. *et al.* The *M*–σ and *M*–*L* relations in galactic bulges, and determinations of their intrinsic scatter. *Astrophys. J.* **698,** 198–221 (2009).
6. Kormendy, J. & Kennicutt, R. C. Secular evolution and the formation of pseudobulges in disk galaxies. *Annu. Rev. Astron. Astrophys.* **42,** 603–683 (2004).
7. Hu, J. The black hole mass–stellar velocity dispersion correlation: bulges versus pseudo-bulges. *Mon. Not. R. Astron. Soc.* **386,** 2242–2252 (2008).
8. Nowak, N. *et al.* Do black hole masses scale with classical bulge luminosities only? The case of the two composite pseudo-bulge galaxies NGC 3368 and NGC 3489. *Mon. Not. R. Astron. Soc.* **403,** 646–672 (2010).
9. Greene, J. E. *et al.* Precise black hole masses from megamaser disks: black hole–bulge relations at low mass. *Astrophys. J.* **721,** 26–45 (2010).
10. Ho, L. C. (ed.) *Coevolution of Black Holes and Galaxies* (Carnegie Observatories Astrophys. Ser. 1, Cambridge Univ. Press, 2004).
11. Silk, J. & Rees, M. J. Quasars and galaxy formation. *Astron. Astrophys.* **331,** L1–L4 (1998).
12. Kormendy, J., Drory, N., Bender, R. & Cornell, M. E. Bulgeless giant galaxies challenge our picture of galaxy formation by hierarchical clustering. *Astrophys. J.* **723,** 54–80 (2010).
13. Toomre, A. in *Evolution of Galaxies and Stellar Populations* (eds Tinsley, B. M. & Larson, R. B.) 401–426 (Yale Univ. Observatory, 1977).
14. Sanders, D. B. *et al.* Ultraluminous infrared galaxies and the origin of quasars. *Astrophys. J.* **325,** 74–91 (1988).
15. Hopkins, P. F. *et al.* A unified, merger-driven model of the origin of starbursts, quasars, the cosmic X-ray background, supermassive black holes, and galaxy spheroids. *Astrophys. J. Suppl. Ser.* **163,** 1–49 (2006).
16. Barth, A. J., Greene, J. E. & Ho, L. C. Dwarf Seyfert 1 nuclei and the low-mass end of the M_{BH}–σ relation. *Astrophys. J.* **619,** L151–L154 (2005).
17. Greene, J. E. & Ho, L. C. The M_{BH}–σ_* relation in local active galaxies. *Astrophys. J.* **641,** L21–L24 (2006).
18. Greene, J. E., Ho, L. C. & Barth, A. J. Black holes in pseudobulges and spheroidals: a change in the black hole–bulge scaling relations at low mass. *Astrophys. J.* **688,** 159–179 (2008).
19. Bentz, M. C., Peterson, B. M., Pogge, R. W. & Vestergaard, M. The black hole mass–bulge luminosity relationship for active galactic nuclei from reverberation mapping and *Hubble Space Telescope* imaging. *Astrophys. J.* **694,** L166–L170 (2009).
20. Woo, J.-H. *et al.* The Lick AGN monitoring project: the M_{BH}–σ_* relation for reverberation-mapped active galaxies. *Astrophys. J.* **716,** 269–280 (2010).
21. Ho, L. C. Nuclear activity in nearby galaxies. *Annu. Rev. Astron. Astrophys.* **46,** 475–539 (2008).
22. Filippenko, A. V. & Ho, L. C. A low-mass central black hole in the bulgeless Seyfert 1 galaxy NGC 4395. *Astrophys. J.* **588,** L13–L16 (2003).
23. Peterson, B. M. *et al.* Multiwavelength monitoring of the dwarf Seyfert 1 galaxy NGC 4395. I. A reverberation-based measurement of the black hole mass. *Astrophys. J.* **623,** 799–808 (2005).
24. Barth, A. J., Strigari, L. E., Bentz, M. C., Greene, J. E. & Ho, L. C. Dynamical constraints on the masses of the nuclear star cluster and black hole in the late-type spiral galaxy NGC 3621. *Astrophys. J.* **690,** 1031–1044 (2009).
25. Thornton, C. E., Barth, A. J., Ho, L. C., Rutledge, R. E. & Greene, J. E. The host galaxy and central engine of the dwarf active galactic nucleus POX 52. *Astrophys. J.* **686,** 892–910 (2008).
26. Hopkins, P. F. & Hernquist, L. Fueling low-level AGN activity through stochastic accretion of cold gas. *Astrophys. J. Suppl. Ser.* **166,** 1–36 (2006).
27. Yu, Q. & Tremaine, S. Observational constraints on growth of massive black holes. *Mon. Not. R. Astron. Soc.* **335,** 965–976 (2002).
28. Schawinski, K. *et al.* Galaxy zoo: the fundamentally different co-evolution of supermassive black holes and their early- and late-type host galaxies. *Astrophys. J.* **711,** 284–302 (2010).
29. Kumar, P. & Johnson, J. L. Supernovae-induced accretion and star formation in the inner kiloparsec of a gaseous disc. *Mon. Not. R. Astron. Soc.* **404,** 2170–2176 (2010).
30. Hopkins, P. F. & Quataert, E. How do massive black holes get their gas? *Mon. Not. R. Astron. Soc.* **407,** 1529–1564 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge with pleasure our collaboration with N. Drory on work leading up to this paper. We thank N. Drory and J. Greene for helpful comments on the manuscript and J. Greene for communicating the maser black-hole detection results before publication. We also thank K. Gebhardt for permission to use $M_\bullet$ values for NGC 4736 and NGC 4826, and J. Jardel for permission to use his updated $M_\bullet$ value for NGC 4594 before publication. Some data used here were obtained with the Hobby–Eberly Telescope (HET), which is a joint project of the University of Texas at Austin, Pennsylvania State University, Stanford University, Ludwig-Maximilians-Universität Munich and Georg-August-Universität Göttingen. It is named in honour of its principal benefactors, W. P. Hobby and R. E. Eberly. We made extensive use of data from the Two Micron All Sky Survey, a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center (IPAC)/California Institute of Technology funded by NASA and by the National Science Foundation (NSF). We also made extensive use of the NASA/IPAC Extragalactic Database (NED), which is operated by California Institute of Technology and the Jet Propulsion Laboratory under contract with NASA; of the HyperLeda database (http://leda.univ-lyon1.fr); and of the NASA Astrophysics Data System bibliographic services. Finally, we are grateful to the NSF for grant support.

Author Contributions J.K. led the programme, carried out the analysis for this paper and wrote most of the text. M.E.C. oversaw the HET observations, preprocessed the HET spectra and provided technical support throughout the project. R.B. calculated the velocity dispersions from the HET spectra and made all least-squares fits. All authors contributed to the writing of the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.K. (kormendy@astro.as.utexas.edu).

LETTER

doi:10.1038/nature09695

Supermassive black holes do not correlate with dark matter haloes of galaxies

John Kormendy[1,2,3] & Ralf Bender[2,3]

Supermassive black holes have been detected in all galaxies that contain bulge components when the galaxies observed were close enough that the searches were feasible. Together with the observation that bigger black holes live in bigger bulges[1–4], this has led to the belief that black-hole growth and bulge formation regulate each other[5]. That is, black holes and bulges coevolve. Therefore, reports[6,7] of a similar correlation between black holes and the dark matter haloes in which visible galaxies are embedded have profound implications. Dark matter is likely to be non-baryonic, so these reports suggest that unknown, exotic physics controls black-hole growth. Here we show, in part on the basis of recent measurements[8] of bulgeless galaxies, that there is almost no correlation between dark matter and parameters that measure black holes unless the galaxy also contains a bulge. We conclude that black holes do not correlate directly with dark matter. They do not correlate with galaxy disks, either[9,10]. Therefore, black holes coevolve only with bulges. This simplifies the puzzle of their coevolution by focusing attention on purely baryonic processes in the galaxy mergers that make bulges[11].

The idea of coevolution was motivated by the observation that bigger black holes live in bulges and elliptical galaxies that have bigger velocity dispersions, σ, at large radii where stars mainly feel each others' gravity and not that of the black hole[3,4]. This correlation was compelling because its scatter was small, consistent with measurement errors. The reduced χ^2 value was 0.79 for the highest-accuracy sample[3], implying that the intrinsic scatter in black-hole mass, $M_\bullet$, at fixed σ is $\lesssim$0.15 dex (ref. 4). The scatter was so small that σ could be used as a surrogate for $M_\bullet$ for many arguments. More important was the implication that a fundamental physical connection between black-hole and bulge growth awaits discovery, especially given the realization[12] that even a tiny fraction of the energy produced in black-hole growth could, if absorbed by protogalactic gas, regulate bulge formation. Small scatter will be important here, too. Tight correlations motivate a search for underlying physics. Loose correlations are less compelling: bigger galaxies just tend to be made of bigger galaxy parts.

The discovery[6,7] of a similarly tight correlation between σ and the circular rotation velocities, $V_{\rm circ}$, of gas in the outer parts of galaxies, where gravity is controlled by dark matter, therefore was taken to imply that dark matter also regulates black-hole growth. In fact, it was suggested[6] that the more fundamental correlation is the one with dark matter; in other words, that dark matter engineers coevolution.

The proposed black hole/dark matter correlation raised two concerns. First, it was known that black holes do not correlate with galaxy disks[9], whereas galaxy disks correlate closely with dark matter[13,14]. It was therefore not clear how black holes and disks could separately correlate with dark matter without also correlating with each other. Second, the velocity resolution of some σ measurements was too low to resolve narrow spectral lines; this problem is discussed in the legend of Fig. 1.

If dark matter controls black-hole growth and bulges are essentially irrelevant, then $V_{\rm circ}$ should correlate tightly with σ even in galaxies that do not have bulges. Figure 1 tests this hypothesis. It also updates the plot that was used to claim[6] a black hole/dark matter correlation. The reliable original data are shown in black; points measured with low velocity resolution were omitted as documented in Supplementary Table 1. Motivated by the above discussion, we measured[8] velocity dispersions in six Sc–Scd galaxies that have nuclear star clusters ('nuclei') but essentially no bulges. They are shown by red points in Fig. 1. Other coloured points show additional published data on bulgeless galaxies that were measured with enough spectral dispersion to resolve nuclear σ.

Bulgeless galaxies (Fig. 1, coloured points plus NGC 3198) show only a weak correlation between $V_{\rm circ}$ and σ. This is expected, because bigger galaxies tend to have bigger nuclei[15]. However, the lack of a tight correlation suggests no more compelling formation physics than the expectation that bigger nuclei can be manufactured in bigger galaxies that contain more fuel. The scatter is much larger than the measurement errors: $\chi^2 = 15.7$. We note in particular that galaxies with $\sigma \approx 25\,{\rm km\,s^{-1}}$ span almost the complete $V_{\rm circ}$ range for the coloured points, 96–210 $\rm km\,s^{-1}$.

Our measurements[8] were made with the 9.2-m Hobby–Eberly Telescope and High Resolution Spectrograph; the instrumental velocity dispersion of $\sigma_{\rm instr} = 8\,{\rm km\,s^{-1}}$ reliably resolves the smallest velocity dispersions seen in galactic nuclei. We confirm that $\sigma = 19.8 \pm 0.7\,{\rm km\,s^{-1}}$ in galaxy M33 and include this value in Fig. 1.

Two properties of our sample deserve emphasis. First, we observed NGC 5457 (also known as M101) and NGC 6946 because these are among the biggest bulgeless galaxies ($V_{\rm circ} \approx 210\,{\rm km\,s^{-1}}$). This is important because it was concluded in ref. 6 that (using our notation) "the $V_{\rm circ}$–σ relation…seems to break down below $V_{\rm circ} \approx 150\,{\rm km\,s^{-1}}$", implying that "halos of mass smaller than $\sim 5 \times 10^{11} M_\odot$ are increasingly less efficient at forming [supermassive black holes]". Our galaxy measurements show that the $V_{\rm circ}$–σ relation breaks down even at $V_{\rm circ} = 210\,{\rm km\,s^{-1}}$ if the galaxy contains no bulge.

Second, our sample is intentionally biased against galaxies that contain bulges. We even avoided substantial 'pseudobulges', that is, 'fake bulges' made by the internal evolution of galaxy disks[16] rather than by the galaxy mergers that make 'classical bulges'. The pseudobulge-to-total mass ratios of our galaxies are a few per cent or less; the bulge-to-total mass ratios are zero. The relevance of pseudobulges is discussed below. We chose these galaxies because, as noted earlier, we want to know whether dark matter correlates with black holes in the absence of the component that we know correlates with black holes. A study[17] of a large galaxy sample that is not biased against bulges results in similar conclusions: $V_{\rm circ}$ correlates weakly with σ, especially for galaxies of Hubble types that commonly contain classical bulges, but the scatter is large and "these results render questionable any attempt to supplant the bulge with the halo as the fundamental determinant of the central black hole mass in galaxies"[17]. Results from this study are included in Supplementary Figure 3.

Figure 1 shows substantial overlap in $V_{\rm circ}$ between the coloured points that show little correlation with σ and the black filled circles that show good correlation. In the overlap range, $180\,{\rm km\,s^{-1}} \lesssim V_{\rm circ} \lesssim 220\,{\rm km\,s^{-1}}$,

[1]Department of Astronomy, University of Texas at Austin, 1 University Station, Austin, Texas 78712-0259, USA. [2]Max-Planck-Institut für Extraterrestrische Physik, Giessenbachstrasse, D-85748 Garching-bei-München, Germany. [3]Universitäts-Sternwarte, Scheinerstrasse 1, D-81679 München, Germany.

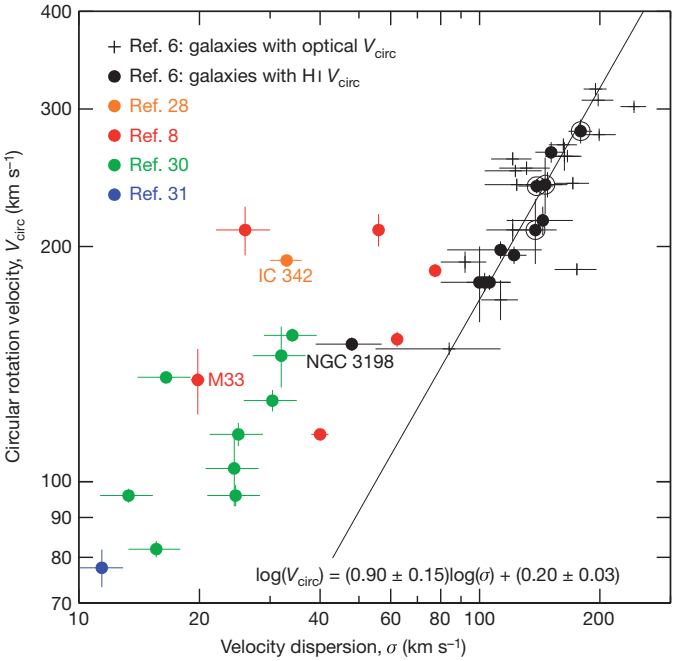


Figure 1 | Outer rotation velocities versus near-central velocity dispersions of disk galaxies. Data are listed in Supplementary Information. Error bars, 1 s.d. The original black hole/dark matter correlation[6] is shown using black symbols (circled if the galaxy has a classical bulge) except that points have been omitted if the σ measurement had insufficient velocity resolution. For example, the bulgeless Scd galaxy IC 342 (now the orange point, after correction) was previously plotted[6] at $\sigma = 77 \pm 12$ km s^{-1}, consistent with the black points. But the measurement[25] had low resolution: the instrumental velocity dispersion, σ_{instr} = (resolution full-width at half-maximum)/2.35, was 61 km s^{-1}, similar to the value of σ measured for IC 342. Low resolution often results in overestimation of σ. The same source lists $\sigma = 77$ km s^{-1} for the nucleus of M33, which has $\sigma = 21 \pm 2$ km s^{-1} as measured at high resolution[26,27]. In fact, a high-resolution measurement of IC 342 was available[28]: at $\sigma_{instr} = 5.5$ km s^{-1}, σ is observed to be 33 ± 3 km s^{-1} (orange point). We replaced four filled circles for which better σ measurements are available; three of these are now coloured points, and the fourth is NGC 3198. We omitted five plus-shaped points for which $\sigma \lesssim \sigma_{instr}$. And we added points (colour) for galaxies measured with $\sigma_{instr} < 10$ km s^{-1}, that is, high enough resolution to allow measurement of the smallest velocity dispersions seen in galactic nuclei. The line (equation at bottom; velocities are in units of 200 km s^{-1}) is a symmetric least-squares fit[29] to the black filled circles minus NGC 3198. It has $\chi^2 = 0.25$. The correlation coefficient is $r = 0.95$. This correlation is at least as good as the one between $M_\bullet$ and σ. The correlation for the plus-shaped points has $\chi^2 = 2.6$ and $r = 0.77$. For these galaxies, we have only optical rotation curves, which measure V_{circ} less accurately than do H I measurements because they do not reach as far out into the dark matter halo[6]. They show a weaker correlation that is not a compelling argument for coevolution. Weaker still is the correlation for the coloured points plus NGC 3198: it has $\chi^2 = 15.7$ and $r = 0.70$.

galaxies satisfy the tight V_{circ}–σ relation shown by the black filled circles only if they contain bulges. We conclude that baryons must matter to black-hole growth. But baryons in a disk are not enough to allow us to predict $M_\bullet$. Dark matter by itself is not enough. The galaxy M101 (Fig. 1, top-left red point) has a halo that is similar to those of half of the tightly correlated galaxies, but that halo did not manufacture a canonical black hole in the absence of a bulge. This suggests that bulges, not haloes, coevolve with black holes.

Nevertheless, most of the black circles in Fig. 1 show a correlation whose scatter is consistent with the error bars. We need to understand this.

We suggest that the tight correlation of black points in Fig. 1 is a result of the well-known 'conspiracy'[13,14] between baryons and dark matter to make featureless rotation curves with no distinction between the parts that are dominated by baryonic and non-baryonic matter. This possibility was considered and dismissed in ref. 6. However, it is a natural consequence of the observation that baryons make up 17% of the matter in galaxies[18] and that, to make stars, they need to dissipate inside their haloes until they are self-gravitating. This is sufficient to engineer that V_{circ} be approximately the same for dark matter haloes and for disks embedded in them[19,20]. That part of the conspiracy is not shown by Fig. 1 because, absent a bulge, disks reach V_{circ} at large radii that are not sampled by σ measurements of nuclei.

Bulges dissipate more than disks. The consequences are shown in Supplementary Figs 1 and 2. Supplementary Fig. 1 shows that V_{circ} for the bulge approximately equals V_{circ} for the halo in the two highest-V_{circ} galaxies whose points are circled in Fig. 1. Supplementary Fig. 2 shows that the same approximate equality holds, given the uncertainties in rotation-curve decomposition, for all decompositions that we could find that included a bulge. It holds in just the V_{circ} range, 180–260 km s^{-1}, where the black circles in Fig. 1 show a tight correlation. Because a bulge has $V_{circ} \approx \sqrt{2}\sigma$, a correlation such as that in Fig. 1 is expected from Supplementary Fig. 2. All galaxies that participate in the tight correlation in Fig. 1 are included in Supplementary Fig. 2, and all of them have bulges or pseudobulges. We conclude that the correlation is nothing more nor less than a restatement of the rotation-curve conspiracy for bulges and dark matter. It is a consequence of dark-matter-mediated galaxy formation. The conceptual leap to a direct causal correlation between dark matter and black holes is not required by the data.

So far, we have discussed black-hole correlations indirectly using the assumption that σ is a surrogate for black-hole mass. We now check this assumption and show that it is not valid for most of the black points that define the tight correlation in Fig. 1. If σ is not a measure of $M_\bullet$ for these galaxies, then this further shows that the correlation is not a consequence of black hole/dark matter coevolution.

In Fig. 2, we examine directly the correlations between $M_\bullet$ and host galaxy properties for galaxies in which black holes have been detected dynamically. All plotted parameters are published elsewhere. The galaxy sample and plotted data are listed in Supplementary Information of the accompanying Letter[10]. The same galaxies are shown in all panels except that ellipticals do not appear in Fig. 2c because they do not have disks; bulgeless galaxies do not appear in Fig. 2a because they do not have bulges; and some bulgeless galaxies and pseudobulge galaxies with $M_\bullet$ limits do not appear in Fig. 2b because σ is outside the range of the plot.

The top panels correlate $M_\bullet$ with the luminosity (Fig. 2a) and velocity dispersion (Fig. 2b) of the host galaxy bulge. Ellipticals (black) and classical bulges (red) show the good (Fig. 2a) and better (Fig. 2b) correlations that we have come to expect.

Figure 2c shows[10] that galaxy disks do not correlate with $M_\bullet$. Disk masses, which are approximately proportional to their K-band luminosities, cannot be used to predict $M_\bullet$.

Figure 2 also distinguishes classical bulges (red points) from pseudobulges (blue points). Classical bulges are essentially indistinguishable in structure and parameter correlations from elliptical galaxies (black points). We believe that both formed by galaxy mergers (see below). Pseudobulges are high-density, central components in galaxies that superficially resemble—and often are mistaken for—classical bulges but that can be recognized because their properties are more disk-like than those of classical bulges. We now know that this results from fundamentally different formation histories. Complementary to hierarchical clustering[21], a new aspect of our understanding of galaxy formation[16] is that isolated galaxy disks evolve slowly as non-axisymmetries such as bars redistribute angular momentum. During this process, pseudobulges are grown out of disk material. Bulge–pseudobulge classifications are listed for all objects in our sample in Supplementary Information of ref. 10. Figure 2a, b illustrates a conclusion from that Letter which has consequences here: pseudobulges show essentially no correlation between $M_\bullet$ and σ. Baryons do not predict $M_\bullet$ if they are in a pseudobulge.

If $M_\bullet$ and σ do not correlate for pseudobulges, then σ is not a surrogate for $M_\bullet$ in Fig. 1, either. Bulge classifications for the galaxies shown in Fig. 1 are given in Supplementary Information, and two

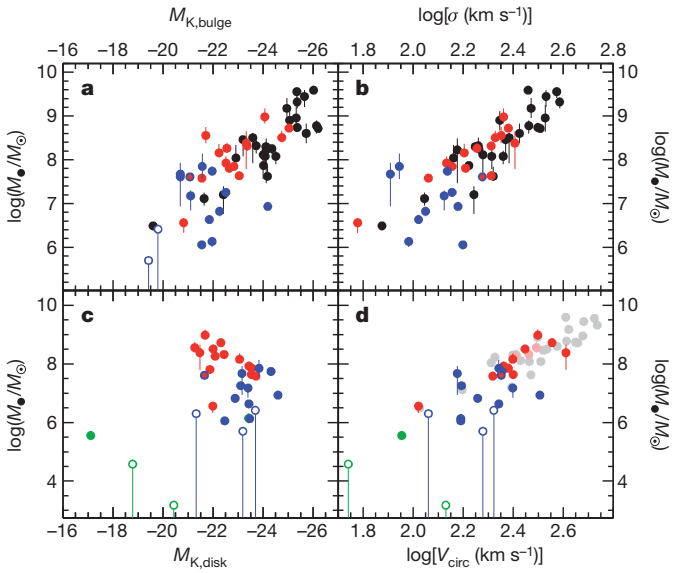


Figure 2 | Correlations of dynamically measured black-hole masses with structural parameters of host galaxies. **a**, Black-hole mass, $M_\bullet$, versus the K-band absolute magnitude of the host galaxy bulge, $M_{K,bulge}$, with disk light removed. **b**, $M_\bullet$ versus the velocity dispersion, σ, of the host bulge averaged inside the radius that contains one-half of the bulge light. Elliptical galaxies are plotted in black, classical bulges are plotted in red and pseudobulges are plotted in blue. The blue symbol with the red centre represents the galaxy NGC 2787, which has a dominant pseudobulge and possibly also a small classical bulge. In least-squares fits, it is included with the pseudobulges. **c**, **d**, Analogous correlations for the disks of host galaxies: $M_\bullet$ versus the K-band absolute magnitude of the disk, $M_{K,disk}$, with bulge light removed (**c**); $M_\bullet$ versus outer rotation velocity, V_{circ} (**d**). Green points are for galaxies that contain neither a classical bulge nor a large pseudobulge but only a nuclear star cluster. Closely similar versions of **a**–**c** appear in the accompanying Letter[10]; all data plotted here are tabulated there in Supplementary Information. Errors bars, 1 s.d. On the basis of these, classical bulges and ellipticals in **a** together have $\chi^2 = 12.1$ and $r = -0.82$; this is the well-known good correlation and is consistent with previous derivations[32]. We assume 1-s.d. errors in M_K of ± 0.05 mag. Similarly, in **b**, the red and black points have $\chi^2 = 5.0$ and $r = 0.89$, implying an intrinsic scatter in log($M_\bullet/M_\odot$) of 0.26 at fixed σ to reduce χ^2 to 1.0. This also is consistent with previous derivations[29,32]. It is the good correlation that motivates our ideas about black hole/bulge coevolution. In contrast, pseudobulges do not correlate with $M_\bullet$; in **a**, the blue points have $\chi^2 = 63$ and $r = 0.27$, and in **b**, they have $\chi^2 = 10.4$ and $r = -0.08$. Similarly[10], disks do not correlate with $M_\bullet$: in **c**, the red and blue points together have $\chi^2 = 81$ and $r = 0.41$. Formally, $M_\bullet$ and disk luminosity anticorrelate, but this is not significant. Finally, in **d**, the blue points have $\chi^2 = 11$ and $r = 0.29$, showing no correlation. In contrast, the red points, for classical bulges, show a correlation that, we argue, is a restatement of the rotation-curve conspiracy. Here elliptical galaxies and some S0 galaxies are plotted in light grey and pink, respectively, to indicate that V_{circ} is not known. We therefore plot the surrogate quantity $\sqrt{2}\sigma$, and it agrees with the correlation seen for bulges that have V_{circ} measurements. **d** is a direct demonstration that, absent bulges, black holes do not correlate with dark matter. In all panels, open symbols represent galaxies for which we have only upper limits on $M_\bullet$; they are not included in the above statistics, but they further strengthen our conclusions.

galaxies with no published data are classified. We find that only four of the black points (circled) in Fig. 1 represent classical bulges; these are M31, NGC 2841, NGC 4258, and NGC 7331. The others represent pseudobulges. For these, the demonstration of a tight V_{circ}–σ correlation is not a demonstration that dark matter and black holes correlate.

Instead, if V_{circ} correlates with σ but σ does not measure $M_\bullet$, then our conclusion that the correlation results from the rotation-curve conspiracy gains further support. Also, the circled points for classical bulges agree with the correlation for pseudobulges. It is implausible that the correlation for the circled points is caused by black hole/dark matter coevolution whereas the identical correlation for the other points has nothing to do with black holes.

Finally, Fig. 2d shows directly that $M_\bullet$ does not correlate with V_{circ} and, therefore, with dark matter for pseudobulges.

An additional argument is given in Supplementary Information, section 3. If dark matter V_{circ} predicts $M_\bullet$ independently of baryon content, then the dark matter with $V_{circ} \approx 1{,}500$ km s^{-1} in clusters of galaxies predicts black holes of mass $M_\bullet \approx 7 \times 10^{11} M_\odot$ that are impossible to hide in well-studied clusters such as Coma.

Therefore, over the whole range of V_{circ} values associated with dark matter, that is, at least 50–2,000 km s^{-1}, Fig. 2d shows a correlation with $M_\bullet$ only for the range 200–400 km s^{-1}, plus NGC 7457 at 105 km s^{-1}, and only if the galaxy contains a classical bulge. The bulge correlation can be understood as an indirect result of the rotation-curve conspiracy. Even for V_{circ} in the range 200–400 km s^{-1}, there is no correlation if the galaxy has only a pseudobulge or disk. Baryons are not irrelevant. They are not even sufficient. To correlate with $M_\bullet$, they must be in a classical bulge or elliptical.

We conclude that black holes do not correlate causally with dark matter haloes. There is no reason to expect that the unknown, exotic physics of non-baryonic dark matter directly affects black-hole growth. Even dark matter gravity is not directly responsible for black hole/galaxy coevolution. Rather, that coevolution seems to be as simple as it could be. Black holes coevolve only with classical bulges and ellipticals. We have a well-developed picture of the their formation. Hierarchical clustering of density fluctuations in cold dark matter results in frequent galaxy mergers[11,21,22]. The products of roughly equal-mass mergers are classical bulges and ellipticals, because progenitor disks get scrambled away by dynamical violence[11]. During this process, gas falls to the centre, triggers a burst of star formation[23,24] and builds the high stellar densities that we see in bulges. This gas may also feed black holes. In fact, we see a correspondence[23] between mergers in progress and quasar-like nuclear activity that increases $M_\bullet$. Our increasingly persuasive picture is that the growth of black holes and the assembly of ellipticals happen together and regulate each other[12,24]. The present results support this picture.

Received 12 July; accepted 19 November 2010.

1. Kormendy, J. in *The Nearest Active Galaxies* (eds Beckman, J., Colina, L. & Netzer, H.) 197–218 (Madrid: Consejo Superior de Investigaciones Científicas, 1993).
2. Kormendy, J. & Richstone, D. Inward bound – the search for supermassive black holes in galactic nuclei. *Annu. Rev. Astron. Astrophys.* **33,** 581–624 (1995).
3. Ferrarese, L. & Merritt, D. A fundamental relation between supermassive black holes and their host galaxies. *Astrophys. J.* **539,** L9–L12 (2000).
4. Gebhardt, K. *et al.* A relationship between nuclear black hole mass and galaxy velocity dispersion. *Astrophys. J.* **539,** L13–L16 (2000).
5. Ho, L. C. (ed.) *Coevolution of Black Holes and Galaxies* (Carnegie Observatories Astrophys. Ser. 1, Cambridge Univ. Press, 2004).
6. Ferrarese, L. Beyond the bulge: a fundamental relation between supermassive black holes and dark matter halos. *Astrophys. J.* **578,** 90–97 (2002).
7. Baes, M., Buyle, P., Hau, G. K. T. & Dejonghe, H. Observational evidence for a connection between supermassive black holes and dark matter haloes. *Mon. Not. R. Astron. Soc.* **341,** L44–L48 (2003).
8. Kormendy, J., Drory, N., Bender, R. & Cornell, M. E. Bulgeless giant galaxies challenge our picture of galaxy formation by hierarchical clustering. *Astrophys. J.* **723,** 54–80 (2010).
9. Kormendy, J. & Gebhardt, K. in *Proc. 20th Texas Symp. Relativ. Astrophys.* (eds Wheeler, J. C. & Martel, H.) 363–381 (American Institute of Physics, 2001).
10. Kormendy, J., Bender, R. & Cornell, M. E. Supermassive black holes do not correlate with galaxy disks or pseudobulges. *Nature* doi:10.1038/nature09694 (this issue).
11. Toomre, A. in *Evolution of Galaxies and Stellar Populations* (eds Tinsley, B. M. & Larson, R. B.) 401–426 (Yale Univ. Observatory, 1977).
12. Silk, J. & Rees, M. J. Quasars and galaxy formation. *Astron. Astrophys.* **331,** L1–L4 (1998).
13. van Albada, T. S. & Sancisi, R. Dark matter in spiral galaxies. *Phil. Trans. R. Soc. Lond. A* **320,** 447–464 (1986).
14. Sancisi, R. & van Albada, T. S. in *Dark Matter in the Universe* (eds Kormendy, J. & Knapp, G. R.) 67–80 (Proc. IAU Symp. 117, Reidel, 1987).
15. Böker, T. *et al.* A Hubble Space Telescope census of nuclear star clusters in late-type spiral galaxies. II. Cluster sizes and structural parameter correlations. *Astron. J.* **127,** 105–118 (2004).
16. Kormendy, J. & Kennicutt, R. C. Secular evolution and the formation of pseudobulges in disk galaxies. *Annu. Rev. Astron. Astrophys.* **42,** 603–683 (2004).
17. Ho, L. C. Bulge and halo kinematics across the Hubble sequence. *Astrophys. J.* **668,** 94–109 (2007).
18. Komatsu, E. *et al.* Five-year Wilkinson Microwave Anisotropy Probe observations: cosmological interpretation. *Astrophys. J. Suppl. Ser.* **180,** 330–376 (2009).

19. Gunn, J. E. in *Dark Matter in the Universe* (eds Kormendy, J. & Knapp, G. R.) 537–546 (Proc. IAU Symp. 117, Reidel, 1987).
20. Ryden, B. S. & Gunn, J. E. Galaxy formation by gravitational collapse. *Astrophys. J.* **318,** 15–31 (1987).
21. White, S. D. M. & Rees, M. J. Core condensation in heavy halos: a two-stage theory for galaxy formation and clustering. *Mon. Not. R. Astron. Soc.* **183,** 341–358 (1978).
22. Springel, V. Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature* **435,** 629–636 (2005).
23. Sanders, D. B. *et al.* Ultraluminous infrared galaxies and the origin of quasars. *Astrophys. J.* **325,** 74–91 (1988).
24. Hopkins, P. F. *et al.* A unified, merger-driven model of the origin of starbursts, quasars, the cosmic X-ray background, supermassive black holes, and galaxy spheroids. *Astrophys. J. Suppl. Ser.* **163,** 1–49 (2006).
25. Terlevich, E., Díaz, A. I. & Terlevich, R. On the behaviour of the IR Ca II triplet in normal and active galaxies. *Mon. Not. R. Astron. Soc.* **242,** 271–284 (1990).
26. Kormendy, J. & McClure, R. D. The nucleus of M 33. *Astron. J.* **105,** 1793–1812 (1993).
27. Gebhardt, K. *et al.* M 33: a galaxy with no supermassive black hole. *Astron. J.* **122,** 2469–2476 (2001).
28. Böker, T., van der Marel, R. P. & Vacca, W. D. CO band head spectroscopy of IC 342: mass and age of the nuclear star cluster. *Astron. J.* **118,** 831–842 (1999).
29. Tremaine, S. *et al.* The slope of the black hole mass versus velocity dispersion correlation. *Astrophys. J.* **574,** 740–753 (2002).
30. Walcher, C. J, *et al.* Masses of star clusters in the nuclei of bulgeless spiral galaxies. *Astrophys. J.* **618,** 237–246 (2005).
31. Ho, L. C. & Filippenko, A. V. High-dispersion spectroscopy of a luminous, young star cluster in NGC 1705: further evidence for present-day formation of globular clusters. *Astrophys. J.* **472,** 600–610 (1996).
32. Gültekin, K. *et al.* The M–σ and M–L relations in galactic bulges, and determinations of their intrinsic scatter. *Astrophys. J.* **698,** 198–221 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Courteau for making available his surface photometry of NGC 801 (Supplementary Information) and J. Greene for helpful comments on the manuscript. The Hobby–Eberly Telescope (HET) is a joint project of the University of Texas at Austin, Pennsylvania State University, Stanford University, Ludwig-Maximilians-Universität Munich and Georg-August-Universität Göttingen. It is named in honour of its principal benefactors, W. P. Hobby and R. E. Eberly. This work was supported by the National Science Foundation.

Author Contributions Both authors contributed to the analysis in this paper. J.K. wrote most of the text.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.K. (kormendy@astro.as.utexas.edu).

LETTER

doi:10.1038/nature09713

Directed self-assembly of a colloidal kagome lattice

Qian Chen[1], Sung Chul Bae[1] & Steve Granick[1,2,3]

A challenging goal in materials chemistry and physics is spontaneously to form intended superstructures from designed building blocks. In fields such as crystal engineering[1] and the design of porous materials[2–4], this typically involves building blocks of organic molecules, sometimes operating together with metallic ions or clusters. The translation of such ideas to nanoparticles and colloidal-sized building blocks would potentially open doors to new materials and new properties[5–7], but the pathways to achieve this goal are still undetermined. Here we show how colloidal spheres can be induced to self-assemble into a complex predetermined colloidal crystal—in this case a colloidal kagome lattice[8–12]—through decoration of their surfaces with a simple pattern of hydrophobic domains. The building blocks are simple micrometre-sized spheres with interactions (electrostatic repulsion in the middle, hydrophobic attraction at the poles, which we call 'triblock Janus') that are also simple, but the self-assembly of the spheres into an open kagome structure contrasts with previously known close-packed periodic arrangements of spheres[13–15]. This open network is of interest for several theoretical reasons[8–10]. With a view to possible enhanced functionality, the resulting lattice structure possesses two families of pores, one that is hydrophobic on the rims of the pores and another that is hydrophilic. This strategy of 'convergent' self-assembly from easily fabricated[16] colloidal building blocks encodes the target supracolloidal architecture, not in localized attractive spots but instead in large redundantly attractive regions, and can be extended to form other supracolloidal networks.

Colloidal crystals are important for their proposed applications in photonics, biomaterials, catalytic supports and lightweight structural materials. They also serve as model systems in which to study the phase behaviour and crystallization kinetics of atomic and molecular crystals[13–15]. Usually composed of hard spheres that are homogeneous in surface functionality, their spontaneous formation is mostly induced by the minimization of entropy, which results in a limited selection of attainable close-packed crystal types[13–15]. More complex crystals assembled from similarly homogeneous spheres have been constructed in binary colloidal[17] and template-assisted systems[18]. To achieve programmable formation of crystals, building blocks with designed specific surface functionalities such as DNA linkers[19,20] and attractive 'patches'[5–7,21,22] have been proposed, but these approaches pose synthetic challenges and can be difficult to generalize. For example, the kagome lattice (see Fig. 1), which is of theoretical interest for mathematical reasons[8] as well as its relevance to mechanical stability of an isostatic lattice[9] and frustration in magnetic materials[10], is composed of interlaced triangles whose vertices have four contacting neighbours. To construct it by direct assembly would require colloids with four unevenly distributed patches on their equators to line up precisely with their counterparts on neighbouring spheres (see Supplementary Fig. 1a) but methods to obtain the desired colloids are not immediately accessible.

Accordingly, we chose the kagome lattice as our target colloidal crystal, and produce it using the following alternative strategy. To reduce the need to start with a specific pattern of attractive spots on each building block, we designed a building block with the orthogonal attributes of minimal surface design combined with self-adjusted coordination number. This simplifies the original four-patch decoration scheme into one with two patches at opposite poles, each of which subtends an angle in the plane large enough to allow coordination with two nearest neighbours (see Supplementary Fig. 1b). This has the advantage that established synthetic methods[16] can be used to decorate spherical particles with two hydrophobic poles of tunable area, separated by an electrically charged middle band. Because each of the hemispheres is chemically 'Janus' (two-sided)[23] with the same middle band, we refer to these as 'triblock Janus'.

This motif causes neighbouring particles to attract at their poles in a geometrical arrangement limited by their size, while avoiding energetically unfavourable contacts between the charged middle bands. After overnight sedimentation, the density mismatch between our gold-plated polystyrene particles and the water in which they are suspended concentrates the particles into a quasi-two-dimensional system. Our synthetic scheme produces elongated caps (see Supplementary Fig. 2), which further facilitates assembly into two-dimensional networks because it allows two nearest neighbours only when the long patch axes of neighbouring particles are parallel. Ordering is then switched on at will by adding salt (3.5 mM NaCl in these experiments) to these spheres in

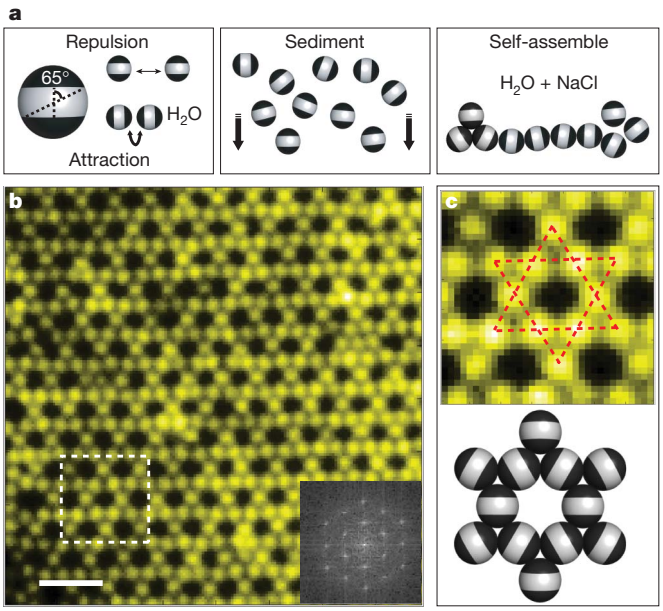


Figure 1 | Colloidal kagome lattice after equilibration. **a**, Triblock Janus spheres hydrophobic on the poles (black, with an opening angle of 65°) and charged in the equator section (white), are allowed to sediment in deionized water. Then NaCl is added to screen electrostatic repulsion, allowing self-assembly by short-range hydrophobic attraction. **b**, Fluorescence image of a colloidal kagome lattice (main image) and its fast Fourier transform image (bottom right). Scale bar is 4 μm. The top panel in **c** shows an enlarged view of the dashed white rectangle in **b**. Dotted red lines in **c** highlight two staggered triangles. The bottom panel in **c** shows a schematic illustration of particle orientations.

[1]Department of Materials Science and Engineering, University of Illinois, Urbana, Illinois 61801, USA. [2]Department of Chemistry, University of Illinois, Urbana, Illinois 61801, USA. [3]Department of Physics, University of Illinois, Urbana, Illinois 61801, USA.

deionized water; the salt screens electrostatic repulsions and allows hydrophobic attraction to come into play. The successful completion of the ordering process requires the energy landscape of the system to be smooth enough for kinetically favoured intermediates to transform finally to the thermodynamically favoured product with maximized hydrophobic contacts. Experimentally, the hydrophobic attraction of around $10k_BT$ (where k_B is the Boltzmann constant and T is temperature) per contact[23] allows self-correction of imperfectly aligned bonds, which is advantageous because kinetically formed intermediate defects finally transform to the favoured structure over a convenient timescale. The scheme of self-assembly is summarized in Fig. 1a.

Fluorescence images of the final product (Fig. 1b) reveal single-phase web-like sheets that tessellate the surface in the known pattern of a kagome lattice. Interestingly, the two populations of pores in this lattice, triangular and hexagonal, each possess inherently different microenvironments on the pore rims. As demonstrated in Fig. 1c, hexagonal cavities are surrounded by negatively charged rims and triangular cavities are surrounded by hydrophobic rims. Previously, others have succeeded in forming kagome lattices when functional molecules self-assemble epitaxially onto certain metals[11], but our substrate is essentially inert; it simply carries negative charge in order to prevent colloids from sticking to it. Accordingly, a key difference from such molecular systems is that lateral positions of elements of the kagome lattice display thermal fluctuation around their mean positions, as shown in Supplementary Movie 1. In principle, this should enable direct measurement of the vibration modes and phonon structure[24], which are of interest theoretically[9]. Although it might seem obvious which structure we will observe, computer simulations of similar particles show a different structure, close-packed with alternating attractive bands[21]. The key difference appears to be that experimentally, particles were free to exchange outside the monolayer. This contrast between experiment and simulation implies that different crystal structures should be observed at low and high pressure. We anticipate another phase transition—melting—if the attraction were weaker, as could be achieved experimentally by using a mixed monolayer of alkane thiols instead of the strongly hydrophobic monolayers used here.

We now follow the ordering process, which is difficult to do with surface-templated systems. Defining the start of the experiment as the moment when salt is added, fluorescence imaging shows that particles first recognize the existence of neighbours by clustering into kinetically favoured triangles, strings or a combination of the two. Subsequently, they rapidly coordinate with additional particles to maximize contacts between hydrophobic poles, resulting in chunks of network with defects consisting mainly of irregular voids. A typical early-stage image is shown in Fig. 2a. The sample at this time contains some web-like structures (Fig. 2b) and other strings with dangling bonds (Fig. 2c). Quantitative analysis of their relative abundance (Fig. 2d) shows, with elapsed time, a monotonic decrease of monomers, a monotonic increase of web-like structures, and a peak in the abundance of strings; hence, strings are a kinetically favoured intermediate that does not require patch alignment as accurate as in webs and that has more translational and rotational freedom. In Supplementary Movie 2, we follow the typical evolution of a string first into triangular nodes with branches, then into enclosed pores. The particle number density in this movie is smaller than in Fig. 2a to demonstrate individual dynamic events. This system behaves similarly to the pattern in Fig. 2d at this stage of assembly, but at a slower rate.

For the kagome lattice to develop fully requires much longer. After the formation of irregular webs, the colloids first adopt quasi-kagome order in local areas, which then extend. This growth of the final structure from metastable intermediates suggests that classical nucleation theory for the (well-studied) crystallization of closely packed spheres should be revisited; in this system, in determining the size and shape of critical nuclei, the concept of average surface tension may be quantitatively or even qualitatively different. To quantify the ordering process,

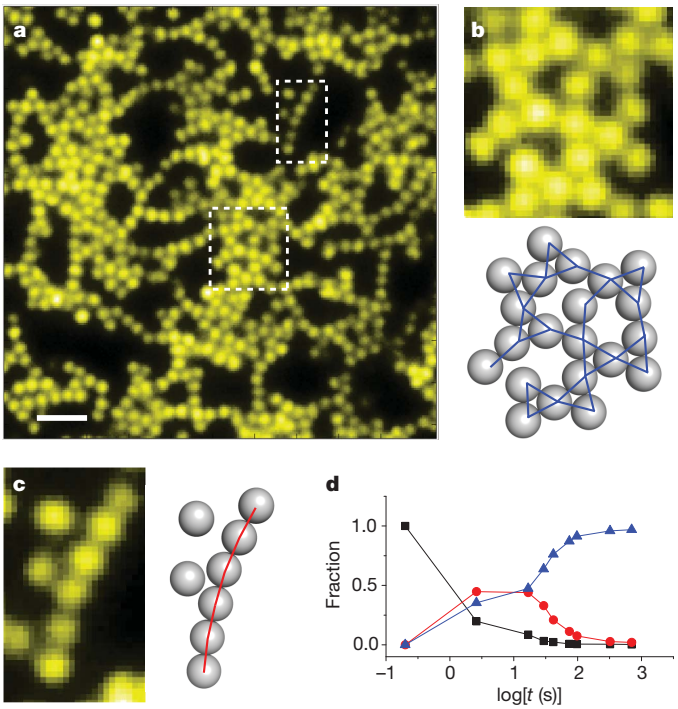


Figure 2 | Stages of self-assembly of kagome lattice. **a**, Illustrative fluorescence image taken 35 s after initiation of assembly. Scale bar is 4 μm. **b**, **c**, Enlarged views of a web-like area (**b**) and a single-particle string area (**c**), both highlighted in **a** by dashed white rectangles. Bonding types are shown in the accompanying schematic illustrations (blue lines in **b** and red lines in **c**). **d**, Time evolution of structures illustrated in **a**–**c**, showing the relative abundance of discrete spheres (black squares), strings (red circles), and triangular bonding (blue triangles). The time spans 0.2 s to 12 min.

for each particle in each time sequence of images, we determined and then quantified the local coordination of each particle to its neighbours using the local bond orientational order parameter ψ_{6j}. Particles with four bonds and ψ_{6j} larger than 0.7 were defined as locally crystalline[14].

Figure 3 consists of a time series of images, showing isolated crystal bonding, then fusion into larger domains, and finally healing of non-crystalline bonds into ordered ones. The crystalline domains fluctuate, making the edges across ordered and random regions rough and sometimes transient (Supplementary Movie 3). The graph in Fig. 3e characterizes the average of images of this kind and shows the time dependence. Interestingly, the approach follows typical first-order chemical reaction kinetics, during which time the particle number density remains nearly constant, with the fraction of product increasing rapidly at first, then more slowly as the available materials are depleted, and finally saturating exponentially to a plateau. Together with the earlier stage in which individual particles condense into irregular networks (Fig. 2 and Supplementary Movie 2), this is reminiscent of two-step nucleation in protein solutions, in which order is preceded by a dense amorphous state[25]. Eventually, we find that growing crystalline grains of different orientations impinge on one another. Supplementary Fig. 3 shows examples of long-range crystalline order, at dilute and concentrated concentrations. This polycrystalline structure could be annealed and its order improved by the usual methods of materials science.

Furthermore, sheets of kagome lattice are found to stack, one above the other, in parallel layers. In a bilayer, both layers retain the in-plane order of a kagome lattice: in registry but staggered in orientation, so that the intersection of nodes of the two lattices forms an octahedron. An optical image of the resulting array of octahedra is shown in Fig. 4. This stacking maximizes the hydrophobic node comprised of six hydrophobic poles in each octahedron; alternative arrangements would be more costly in energy, requiring hydrophobic nodes in one plane to be positioned over the pores with charged rims in the second

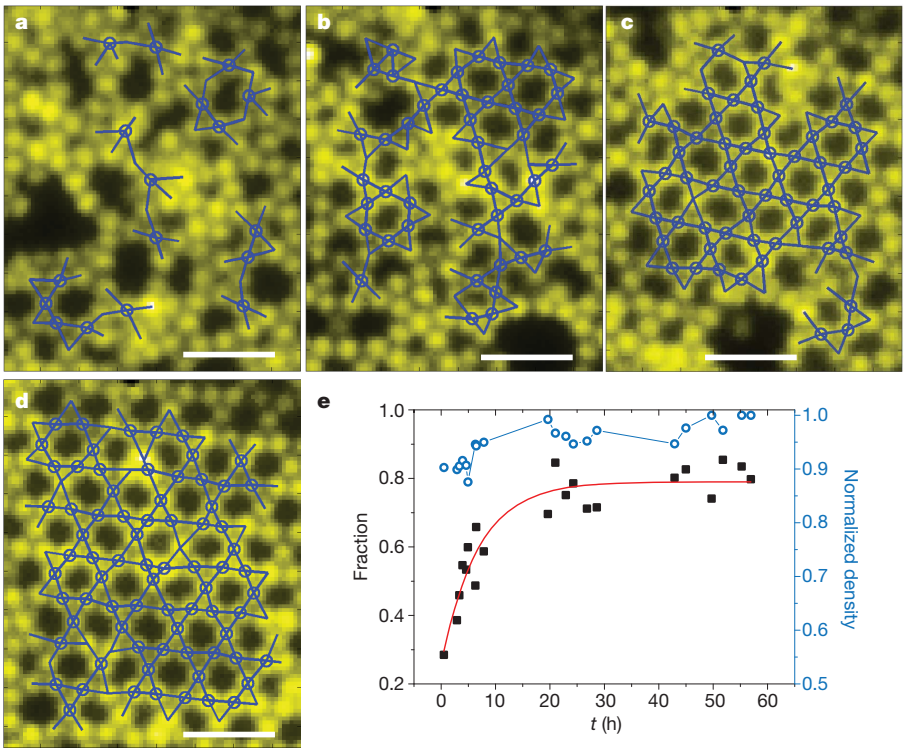


Figure 3 | Crystallization of the kagome lattice. Fluorescence images taken at 2.9 h, 3.0 h, 3.1 h and 53.8 h (**a**, **b**, **c** and **d**, respectively). Blue circles denote the particles of local crystalline order as defined in the text. Blue lines are their neighbouring bonds. Scale bar is 4 μm. **e**, The time evolution of the fraction of particles bearing local crystalline order (black squares), corrected for particles with missing bonds at the boundary, which approaches a plateau with exponential kinetics (red line). Also shown is the particle number density (blue circles), normalized by its observed maximum.

plane. Although technical limitations in visualizing the resulting structure at present prevent us from visualizing the assembly to still thicker films, there is the potential to fabricate multilayers of vertex-sharing octahedral, pyrochlore and other hierarchical structures usually associated with inorganic crystals rather than colloids[8]. Looking ahead to possible applications, freezing these structures into place offers the potential to form selective membranes, in which some holes are hydrophobic and others hydrophilic.

These design rules suggest generalizations. Other open structures[26] could be designed from triblock particles the coordination numbers of which differ on the opposing north and south poles. The needed modulation of the angular range of attraction could be achieved by lessening the size of the hydrophobic patch, or alternatively by less screening of repulsion. If building blocks were to carry four attractive patches distributed at tetrahedral angles, a diamond structure would be artificially designed, although implementation of this awaits the development of the synthetic methods needed to produce the parent particles[27]. The common point is that colloidal building blocks, attracting one another reversibly, during the early stages of assembly assemble into intermediate clusters (strings, in the present experiments) with wide latitude in the mutual orientation of neighbouring particles. Subsequently, the orthogonal variable of geometrical shape then guides these transiently stable intermediates to the final structure.

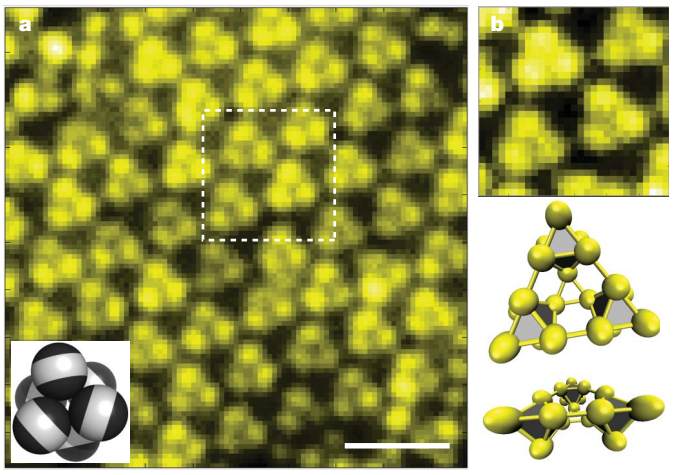


Figure 4 | A bilayer of parallel kagome lattices. **a**, Fluorescence image taken from the bottom. The eye sees octahedra consisting of staggered triangular nodes, as shown in the schematic illustration at bottom left. Scale bar is 4 μm. **b**, Enlarged view of the dashed white square in **a**. The top panel in **b** shows a fluorescence image. The bottom panel in **b** shows a schematic view of particle arrangements from two perspectives, nearly vertical to the plane and nearly parallel to it.

METHODS SUMMARY

Fluorescent latex particles of sulphate polystyrene (1 μm in diameter, F-8851 from Invitrogen) are made hydrophobic on opposite poles through glancing angle deposition of titanium (2 nm) and gold (25 nm) thin films, followed by the deposition of self-assembled monolayers of *n*-octadecanethiol (Sigma-Aldrich). After adding salt (NaCl) to deionized water to a final concentration of 3.5 mM, their self-assembly is observed at room temperature under epifluoresence microscopy. Images captured by an iXon electron multiplying charge coupled device (EMCCD) camera are analysed manually for the early stage of self-assembly, and by single-particle tracking code after particles begin to form web-like structures. Because incomplete webs are only locally periodic, we use the local bond orientational order parameter, rather than a translational order parameter, to quantify the growth of ordered regions within the sample.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 22 July; accepted 25 November 2010.

1. Desiraju, G. R. Crystal engineering: a holistic view. *Angew. Chem. Int. Edn Engl.* **46,** 8342–8356 (2007).
2. Bartels, L. Tailoring molecular layers at metal surfaces. *Nature Chem.* **2,** 87–95 (2010).

3. Yaghi, O. M. *et al.* Reticular synthesis and the design of new materials. *Nature* **423,** 705–714 (2003).
4. Sun, Q.-F. *et al.* Self-assembled $M_{24}L_{48}$ polyhedra and their sharp structural switch upon subtle ligand variation. *Science* **328,** 1144–1147 (2010).
5. Glotzer, S. C. & Solomon, M. J. Anisotropy of building blocks and their assembly into complex structures. *Nature Mater.* **6,** 557–562 (2007).
6. Mann, S. Self-assembly and transformation of hybrid nano-objects and nanostructures under equilibrium and non-equilibrium conditions. *Nature Mater.* **8,** 781–792 (2009).
7. Liu, K. *et al.* Step-growth polymerization of inorganic nanoparticles. *Science* **329,** 197–200 (2010).
8. van der Marck, S. C. Site percolation and random walks on *d*-dimensional Kagomé lattices. *J. Phys. Math. Gen.* **31,** 3449–3460 (1998).
9. Souslov, A., Liu, A. J. & Lubensky, T. C. Elasticity and response in nearly isostatic periodic lattices. *Phys. Rev. Lett.* **103,** 205503 (2009).
10. Atwood, J. L. Kagomé lattice: a molecular toolkit for magnetism. *Nature Mater.* **1,** 91–92 (2002).
11. Schlickum, U. *et al.* Chiral Kagomé lattice from simple ditopic molecular bricks. *J. Am. Chem. Soc.* **130,** 11778–11782 (2008).
12. Glettner, B. *et al.* Liquid-crystalline Kagome. *Angew. Chem. Int. Edn Engl.* **47,** 9063–9066 (2008).
13. Yethiraj, A. & van Blaaderen, A. A colloidal model system with an interaction tunable from hard sphere to soft and dipolar. *Nature* **421,** 513–517 (2003).
14. Gasser, U., Weeks, E. R., Schofield, A., Pusey, P. N. & Weitz, D. A. Real-space imaging of nucleation and growth in colloidal crystallization. *Science* **292,** 258–262 (2001).
15. Anderson, V. J. & Lekkerkerker, H. N. W. Insights into phase transition kinetics from colloid science. *Nature* **416,** 811–815 (2002).
16. Pawar, A. B. & Kretzschmar, I. Patchy particles by glancing angle deposition. *Langmuir* **24,** 355–358 (2008).
17. Shevchenko, E. V., Talapin, D. V., Kotov, N. A., O'Brien, S. & Murray, C. B. Structural diversity in binary nanoparticle superlattices. *Nature* **439,** 55–59 (2006).
18. Xia, Y., Yin, Y., Lu, Y. & McLellan, J. Template-assisted self-assembly of spherical colloids into complex and controllable structures. *Adv. Funct. Mater.* **13,** 907–918 (2003).
19. Park, S. Y. *et al.* DNA-programmable nanoparticle crystallization. *Nature* **451,** 553–556 (2008).
20. Nykypanchuk, D., Maye, M. M., van der Lelie, D. & Gang, O. DNA-guided crystallization of colloidal nanoparticles. *Nature* **451,** 549–552 (2008).
21. Giacometti, A., Lado, F., Largo, J., Pastore, G. & Sciortino, F. Effects of patch size and number within a simple model of patchy colloids. *J. Chem. Phys.* **132,** 174110 (2010).
22. Doppelbauer, G., Bianchi, E. & Kahl, G. Self-assembly scenarios of patchy colloidal particles in two dimensions. *J. Phys. Condens. Matter* **22,** 104105 (2010).
23. Hong, L., Cacciuto, A., Luijten, E. & Granick, S. Clusters of amphiphilic colloidal spheres. *Langmuir* **24,** 621–625 (2008).
24. Ghosh, A., Chikkadi, V. K., Schall, P., Kurchan, J. & Bonn, D. Density of states of colloidal glasses. *Phys. Rev. Lett.* **104,** 248305 (2010).
25. Liu, H., Kumar, S. K. & Douglas, J. F. Self-assembly-induced protein crystallization. *Phys. Rev. Lett.* **103,** 018101 (2009).
26. Grünbaum, B. & Shephard, G. C. *Tilings and Patterns* (W. H. Freeman, 1987).
27. Kraft, D. J., Groenewold, J. & Kegel, W. K. Colloidal molecules with well-controlled bond angles. *Soft Matter* **5,** 3823–3826 (2009).
28. Li, C., Hong, G. & Qi, L. Nanosphere lithography at the gas/liquid interface: a general approach toward free-standing high-quality nanonets. *Chem. Mater.* **22,** 476–481 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the US Department of Energy, Division of Materials Science, under award number DE-FG02-07ER46471 through the Frederick Seitz Materials Research Laboratory at the University of Illinois at Urbana-Champaign. For equipment, we acknowledge the National Science Foundation, CBET-0853737. We thank K. Chen for help with particle tracking.

Author Contributions Q.C. and S.G. initiated this work; Q.C. and S.G. designed the research programme; Q.C. performed the experiments; Q.C., S.C.B. and S.G. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.G. (sgranick@illinois.edu).

METHODS

Particle fabrication. Patches on opposite poles of spherical particles are produced by sequential glancing angle deposition[16]. First, a closely packed monolayer of 1 μm fluorescent sulphated latex particles (F-8851, Invitrogen) is fabricated on a silicon wafer substrate using a reported method[28]. In brief, 80 μl water/ethanol dispersion (volume ratio 1:1) containing 8 wt% latex particles is dropped onto the top surface of a 1 cm × 1 cm piece of glass (pretreated by Piranha solution) surrounded by water located at the midbottom of a Petri dish. The dispersion spreads freely on the water surface until it covers nearly the entire area. Then 10 μl of sodium dodecyl sulphate (2 wt%) solution is added to the water surface to reduce the surface tension and condense the particles into a closely packed monolayer of about 16 cm^2 in area. A silicon wafer (1.5 cm × 2.5 cm, pretreated by Piranha solution) is used to pick a piece of floating monolayer of particles, left to dry for later treatment.

Second, glancing angle deposition of 2 nm Ti/25 nm Au layers onto the colloidal monolayer is performed as described[16]. The glancing angle θ is set to be 30° to the particle monolayer. After the first vapour deposition, the particle monolayer is lifted up with a polydimethylsiloxane (PDMS) stamp so that patches from the first vapour deposition are facing down. PDMS stamps are prepared by curing the monomer and crosslinking agent (10:1 w/w) (Dow Corning) at 70 °C in a pumped oven overnight. Just before stamping, the PDMS surface is treated with oxygen plasma to induce the necessary adhesion. The oxygen plasma is generated by a Harrick PDC-32G plasma cleaner. Low plasma power is used (6.8 W), and the chamber pressure is about 150 mTorr. The treatment duration is 45 s. The stamping is carried out immediately after the plasma treatment. Then the second deposition is performed, from the other direction of the colloidal monolayer, to produce patches on the other poles of the colloids. The PDMS stamp with colloidal particles attached is immersed in 2 mM octadecanethiol (Sigma-Aldrich) in ethanol for 7 h to render the Au coatings hydrophobic. Particles on PDMS stamp are rinsed with ethanol multiple times and then redispersed in deionized water via ultrasonication.

Self-assembly. A suspension of triblock Janus particles in deionized water is contained in a flat silica cuvette (Lab-Tek II chambered coverglass). Particles are repelled from the cuvette bottom by its negative charge. We note that the gravitational height of the as-prepared particles, $h = k_B T/mg \approx 4$ μm, concentrates the dispersed particles into a quasi-two-dimensional system after overnight sedimentation. Here, T is the room temperature and mg is the buoyant weight of the particle, considering both the density of latex particles (1.055 g cm^{-3}) and the gold coating. The range of hydrophobic attraction is short compared to our particle diameter, which the experimental literature reports to be in the range 10–100 nm and was successfully modelled with a potential decaying roughly exponentially with a decay constant of the order of 10 nm (ref. 23). This short range can enforce contact interactions, disfavouring the more loosely bound structures that will result for long-range attractions. The salt NaCl is added in order to screen repulsion to a Debye length of ~5 nm to allow the recognition of attraction between patches and their consequent self-assembly. Using epifluorescence microscopy (63× air objective with a 1.6× post magnification, numerical aperture 0.75) with an iXon EMCCD camera, we monitor the dynamic evolution of the system. We quantified this dynamical change in structure by manual mapping in Fig. 2d, and a combination of particle tracking and calculation of the local two-dimensional bond-orientational order parameter in Fig. 3e:

$$\psi_{6j} = \frac{1}{nn}\sum_{k=1}^{nn} e^{6i\theta_{jk}}$$

where nn is the number of nearest neighbours of particle j identified from Delaunay triangulation. Here θ_{jk} is the angle of the bond between particle j and its neighbour k to an arbitrary reference axis. This definition of order parameter is valid for the kagome lattice because it shares the same arrangement of neighbour orientation as a triangular lattice except for the missing particles. Meanwhile, to avoid the inclusion of particles locally ordered with six bonds as in a triangular lattice, only particles with both $\psi_{6j} > 0.7$ and four nearest-neighbouring bonds are denoted as particles with local kagome lattice order.

LETTER

doi:10.1038/nature09698

Probing the electromagnetic field of a 15-nanometre hotspot by single molecule imaging

Hu Cang[1,2], Anna Labno[2,3], Changgui Lu[2], Xiaobo Yin[1,2], Ming Liu[2], Christopher Gladden[2], Yongmin Liu[2] & Xiang Zhang[1,2]

When light illuminates a rough metallic surface, hotspots can appear, where the light is concentrated on the nanometre scale, producing an intense electromagnetic field. This phenomenon, called the surface enhancement effect[1,2], has a broad range of potential applications, such as the detection of weak chemical signals. Hotspots are believed to be associated with localized electromagnetic modes[3,4], caused by the randomness of the surface texture. Probing the electromagnetic field of the hotspots would offer much insight towards uncovering the mechanism generating the enhancement; however, it requires a spatial resolution of 1–2 nm, which has been a long-standing challenge in optics. The resolution of an optical microscope is limited to about half the wavelength of the incident light, approximately 200–300 nm. Although current state-of-the-art techniques, including near-field scanning optical microscopy[5], electron energy-loss spectroscopy[6], cathode luminescence imaging[7] and two-photon photoemission imaging[8] have subwavelength resolution, they either introduce a non-negligible amount of perturbation, complicating interpretation of the data, or operate only in a vacuum. As a result, after more than 30 years since the discovery of the surface enhancement effect[9–11], how the local field is distributed remains unknown. Here we present a technique that uses Brownian motion of single molecules to probe the local field. It enables two-dimensional imaging of the fluorescence enhancement profile of single hotspots on the surfaces of aluminium thin films and silver nanoparticle clusters, with accuracy down to 1.2 nm. Strong fluorescence enhancements, up to 54 and 136 times respectively, are observed in those two systems. This strong enhancement indicates that the local field, which decays exponentially from the peak of a hotspot, dominates the fluorescence enhancement profile.

The study of the surface enhancement effect mirrors the development of surface analysis techniques. Previous optical experiments revealed that the roughness of the surface has a critical role in determining the strength of the enhancement[9,11]. Further experimental[12] and theoretical[13] studies on the impact of the surface roughness led to the connection between the surface enhancement effect and the surface plasmon, as well as to the hotspots being termed localized surface plasmon polaritons[4,13–15]. More recently, near-field scanning optical microscopy[16], electron energy-loss spectroscopy[6], cathodoluminescence imaging[7], two-photon photoemission imaging[8] and experiments with a waveguide mode excitation[17] have shown that, at these hotspots, the fluorescence enhancement is confined to a region far smaller than the wavelength of light, yet the field of a single hotspot has not been resolved.

We developed a single molecule super-resolution imaging method, based on the fact that individual fluorescent molecules can be localized with single-nanometre accuracy from the optical far field. When multiple emitters reside within a diffraction-limited spot, one has to ensure that these emitters emit one at a time[18–21]. Control of the emitting sequence can be done by serial photo-switching of fluorescence molecules[18,19], using techniques such as photo-activated localization microscopy[18] and stochastic optical reconstruction microscopy[19]. These techniques, together with structural illumination microscopy and stimulated emission depletion microscopy, allow optical resolution beyond the diffraction limit in biological samples (for a review, see ref. 22). As photo-switching would be infeasible for investigating local field, we use the Brownian motion of single dye molecules[20,21,23] in a solution to let the dyes scan the surface of single hotspots in a stochastic manner, one molecule at a time.

Figure 1a illustrates the principle of this new technique. Here, a sample is submerged in a solution of fluorescent dye. The chamber containing the sample is mounted on a total internal reflection (TIRF) set-up. As the diffusion of the dye molecules is much faster than the image acquisition time (a 1-nm-diameter sphere diffuses through a 200-nm-wide spot in less than 0.1 ms on average in water at room temperature, in contrast to an imaging time of typically 50–100 ms), the fluorescence from the rapidly diffusing dye molecules contributes only to a homogeneous background. When a dye molecule is adsorbed onto the surface of a hotspot, it appears as a bright spot. By using a maximum likelihood single molecule localization method[24] (Supplementary Information), the molecule can be localized with single-nanometre accuracy[18–21], and the fluorescence enhancement can be deduced from the intensity of the fluorescence. After the dye molecule is bleached, which typically occurs within hundreds of milliseconds, the fluorescence disappears and the hotspot is ready for the next adsorption event. By choosing the right concentration of dye molecules, the adsorption rate can be controlled to the point of ensuring that only one molecule emits at a time. As we use a camera to record single molecule adsorption events, multiple hotspots within a large field of view of up to 1 mm^2 can be imaged in parallel. This is much more efficient than raster scanning based techniques, such as near-field scanning optical microscopy and electron energy-loss spectroscopy.

Using this method, we are able to image the fluorescence enhancement profile of single hotspots as small as 15 nm with an accuracy down to 1.2 nm, well beyond the diffraction limit, in just a few minutes (Supplementary Movie; Fig. 1b). The hotspots are formed on the surface of a thin (12–15 nm) aluminium film deposited on a quartz substrate by electron-beam evaporation. Water facilitates the oxidization process[25]; an oxidized layer can form on the surface of the aluminium film and reach a thickness of up to 8 nm (ref. 25). Each sphere in Fig. 1b represents a single molecule adsorption event, with the x and y coordinates representing the location of the molecule's centroid, and the z coordinate corresponding to the intensity of the molecule's fluorescence. Single molecules can be localized with an accuracy down to 1.2 nm (Fig. 1c; Supplementary Information). The measured enhancement here exhibits a rapid exponential decay with a decay constant of 9.8 nm (Fig. 1d, e). The standard deviation of the location of observed single molecule events was used as a model-independent measure of the width (15.4 nm) of the hotspot, as shown in Fig. 1f. The existence of such a small hotspot demonstrates the presence of tight near-field optical confinement.

The accurate estimation of the hotspot enhancement and electromagnetic field experienced by the molecule is confounded by a number

[1]Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [2]NSF Nano Scale Science and Engineering Center (NSEC), 3112 Etcheverry Hall, University of California, Berkeley, California 94720, USA. [3]Biophysics Program, University of California Berkeley, Berkeley, California 94720, USA.

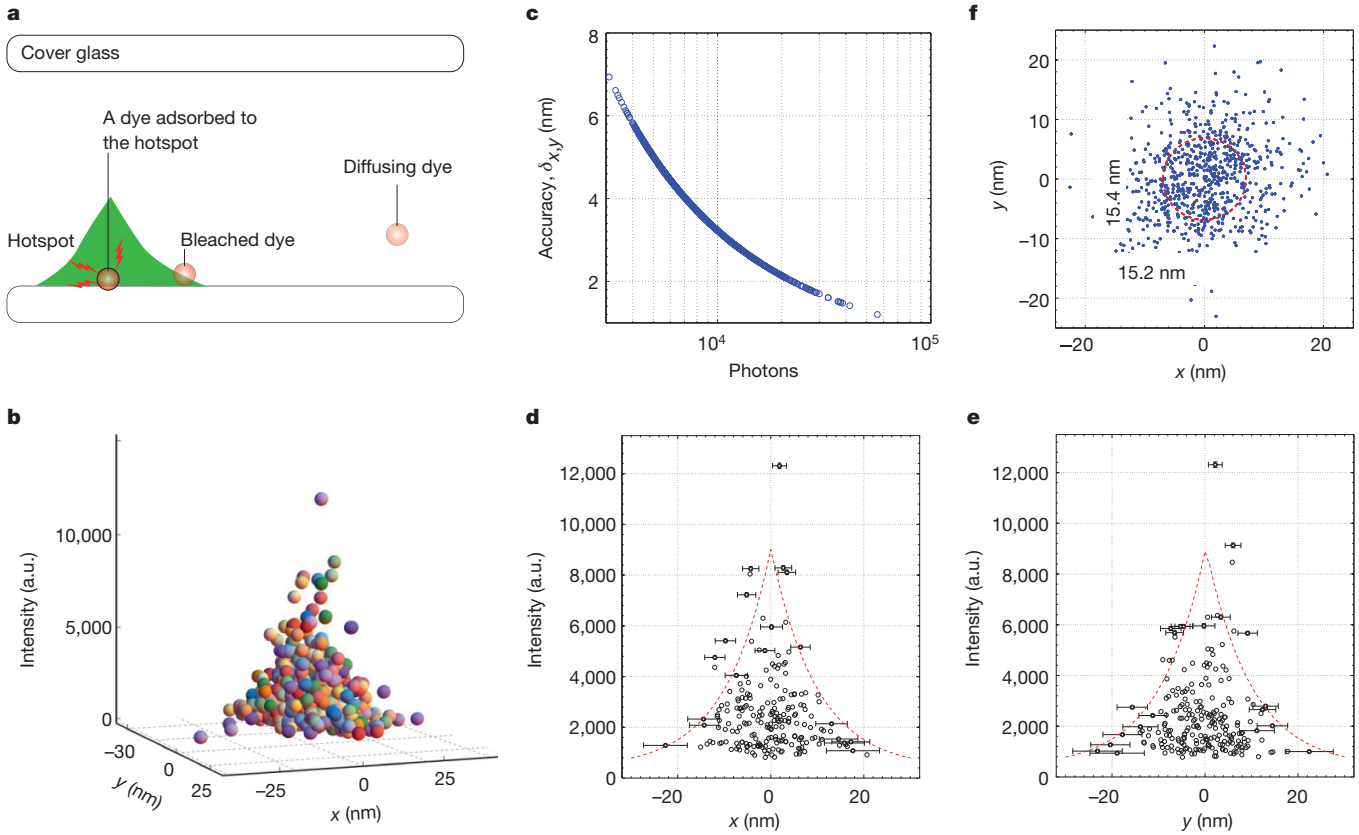


Figure 1 | The principle of Brownian motion single molecule super-resolution imaging. **a**, Hotspots appear on the surface of a thin aluminium film under a total internal reflection (TIRF)-type illumination at 532 nm. To map the field distribution inside the hotspots, we use the Brownian motion of fluorescence dye molecules (Chromeo 546). The dye molecules stochastically adsorb to the surface. After a few frames (50–100 ms per frame), the dye molecules photobleach, give rise to a 'blinking' pattern (Supplementary Movie): each blink corresponds to one adsorption-bleaching event. By controlling the concentration of the dyes, the adsorption rate can be adjusted to ensure that within a diffraction limited spot, only one molecule emits photons at a time; therefore the position of the molecule can be determined by a maximum likelihood localization method[24] with accuracy down to 1.2 nm (**c**). **b**, By using the adsorption locations as the x and y coordinates, and the fluorescence intensity as the z coordinate, we obtain a 3D scatter plot of the fluorescence enhancement profile of the hotspot, with each sphere representing one single molecule event. **c**, The accuracy of the reconstructed field profile, estimated from the variance of the maximum likelihood localization[24] (Supplementary Information), depends on the number of photons collected from the molecules, with brighter single molecule events showing better accuracy, down to 1.2 nm. The molecules within $-2\ \text{nm} < y < 2\ \text{nm}$ are shown in **d**, which represents a cross-section of the hotspot at $y = 0$ nm. A similar plot of the cross-section of the hotspot at $x = 0$ nm is shown in **e**. The envelope appears as an exponential decay with a constant of 9.83 nm (red dashed lines). To avoid crowding, only the single molecule localization variance of a few spheres near the envelope is shown. **f**, The distribution of the single molecule events provides a direct measure of the size of the hotspot; the width of the hotspot characterized by the standard deviation of the single molecule events is 15.2 nm and 15.4 nm in x and y directions, respectively.

of factors, including the molecule's orientation relative to the polarization of local field, and stochastic photobleaching (Supplementary Information). To decrease the effect of those stochastic variables on the final shape of the profile, we use a Gaussian kernel method to render the image of a hotspot[18], as shown in Fig. 2a. Each pixel X of the rendered image, $I(X)$, is a weighted average of the intensity from all of the single molecule events, with molecules closer to X carrying more weight (Fig. 2a). The window size of the kernel is determined by the accuracy of the single molecule localization. The rendered image reports an averaged profile of the local fluorescence enhancement. A hotspot may have complex structures that are smaller than the size of the kernel; these fine structures are removed in the stochastic rendering process (Supplementary Information). The exponential decay is evident, with the full-width at half-maximum (FWHM) measured as ~20 nm (Fig. 2b).

By comparing the fluorescence from the dye molecules adsorbed on the hotspot to that from the dye molecules immobilized on the surface of a quartz slide under the same conditions (Supplementary Information), we determined that the fluorescence enhancement is about 36 times greater at the centre of the 20-nm hotspot. This modest enhancement results from the high ohmic loss of aluminium at the visible wavelength involved. Among the total of 60 hotspots that were analysed, we found a broad distribution of enhancement factors and sizes, with approximately 32 nm as the average size of the spots (Fig. 2c). The maximum fluorescence enhancement factor of the hotspots was found to depend inversely on their size; the largest enhancement factor we observed was 54 times (for a 15-nm hotspot); this results from the tighter confinement of the electromagnetic field in a smaller hotspot (Fig. 2d).

As hotspots also appear in metal nanoparticle clusters[12,13], and the mechanism has been postulated to be similar to that of the metal thin films, we investigated the hotspots that formed in silver clusters consisting of silver nanoparticles with average diameter of 40 nm. In addition to the oxidation layer of silver[26], the dispersing surfactant (that comes with the silver nanoparticle suspension) condenses on the surface of the silver nanoparticle clusters during the aggregation process. We found that the fluorescence enhancement profile of the hotspots formed in the silver clusters has an exponential shape, similar to that of hotspots on the surface of an aluminium film (Fig. 3a, b). The electromagnetic field is strongly confined in a hotspot within an elliptical region of 13.2 nm × 20.3 nm, more than 30 times smaller than the wavelength of the excitation laser (Fig. 3c).

The fluorescence enhancement includes contributions from both local field and quenching. Experiments have shown that the quenching—transfer of energy from molecules to the metal nanostructure[27]—is

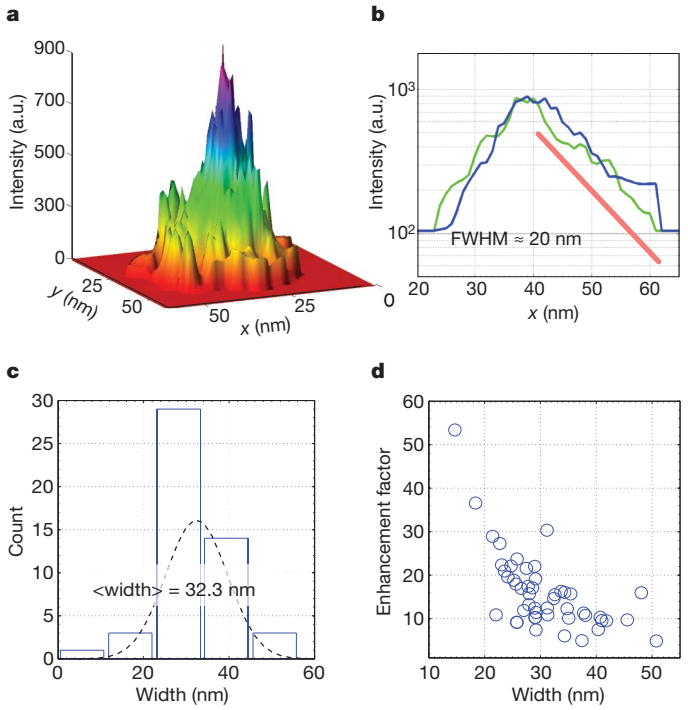


Figure 2 | Hotspots on an aluminium film. **a**, As the fluorescence of a single molecule is intrinsically stochastic, we remove this randomness by using a Gaussian kernel method to render the image of the field distribution. A hotspot on the surface of an aluminium film is shown. Each pixel *X* of the rendered image corresponds to average intensity from all of the single molecule events, with molecules closer to *X* carrying more weight. The kernel window size is 2.1 nm; this small window size makes the image appear noisy. An exponential decay field profile is visible, and is more evident on a log scale, shown in **b** as almost a decade of straight line (red solid line). The blue and green curves are two cross-sections of the hotspot along *x* and *y* directions through the peak. The FWHM of the spot is ~20 nm. **c**, All of the hotspots observed are of deep sub-wavelength size, with an average width of 32.3 nm. By plotting the enhancement factor of the hotspots against their width (**d**), an inverse relationship between the size and the enhancement factor is visible: tighter confinement leads to stronger enhancement.

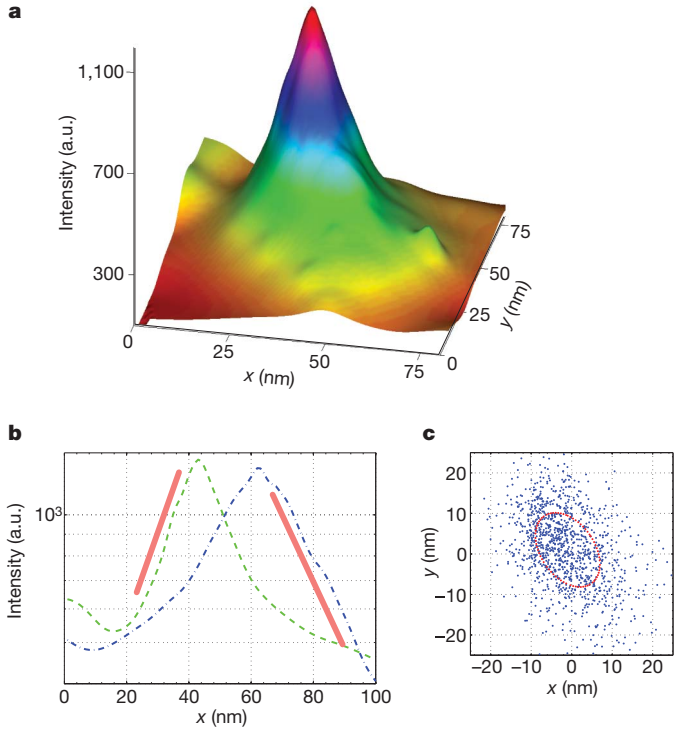


Figure 3 | A hotspot on silver nanoparticle clusters. **a**, A hotspot formed on silver nanoparticle clusters appears similar to those formed on the aluminium film. A 644-nm laser is used for excitation, and Chromeo 642 dye (Active Motif)—whose emission centres around 660 nm—is used. The maximum enhancement factor at the centre of the peak corresponds to 136 times the florescence from the same dye molecules adsorbed on a glass surface. **b**, The hotspot exhibits a similar exponential decay profile as those formed on the surface of the aluminium film. Two cross-sections of the hotspot along *x* (green) and *y* (blue) directions through the peak are plotted on a logarithmic scale, with the solid red lines as eye-guides for the exponential profile. **c**, The widths of the hotspot, estimated from the distribution of the single molecule events on a scatter plot, are 13.2 nm and 20.3 nm along the two axes.

short-ranged compared to the local field. Therefore there exist two distinct regimes above the surface of a metal: a quenching dominated regime near the surface, followed by a local field dominated regime further away from the surface[26,28]; the transition point between the two regimes coincides with the peak of the fluorescence enhancement. This peak ranges from a few nanometres to tens of nanometres, depending on the materials and the geometry of the nanostructures[26,28]. The dielectric layers formed on the surface of the metal shift the adsorbed layer of molecules into the local field dominant regime, contributing to the strong (up to 54 times (Al) and 136 times (Ag)) fluorescence enhancement observed. To reproduce the effect of the dielectric layers, we deposited a ~9-nm-thick spacer layer of self-assembled protein molecules on the surface of silver nanoparticle clusters *in situ*, in addition to the original dielectric layer (Supplementary Information). As the transition point of silver nanoparticle clusters is ~2.5 nm above the surface in a similar system[26], the observation will be in the local field dominant regime. Comparing the same hotspot without and with the spacer layer, we found an ~36% drop in the fluorescence enhancement, an ~26% increase in width, and an essentially unchanged exponential shape (Supplementary Information), which confirms that an exponential profile is a general feature of the local electromagnetic field.

The single molecule super-resolution approach is a generic technique, which offers a unique, perturbation free capability for imaging the electromagnetic field enhancement of optical nanostructures with single-nanometre precision. Using this technique, we have demonstrated the first (to our knowledge) direct measurement of a single hotspot as small as 15 nm and with an accuracy down to 1.2 nm. Such measurements will accelerate understanding of localized electromagnetic modes. The exponential profile observed sheds new light on the much-debated mechanism of extraordinary field confinement in a two-dimensional disordered system where Anderson-localized modes[4,15] could emerge. Although signatures of the Anderson-localized mode have been reported in two-dimensional disordered metallic systems[16], its hallmark, an exponential profile, had not been directly observed up until now. Additionally, other imaging modalities can be integrated to provide spectroscopy as well as fluorescence lifetime information on the molecules; this approach could be used to investigate the strong-coupling regime such as occurs with resonant nano-antennas[29,30]. In the strong-coupling regime, the distributed dipoles in the antenna could significantly affect the fluorescence of the molecules, and the strength of the coupling cannot be directly measured from the intensity of the fluorescence signals. A super-resolution measurement of the spatial distribution of the molecules' fluorescence lifetime will help visualization and understanding of the coupling. Furthermore, with the development of brighter fluorophores, better photo-bleaching suppression methods, and more efficient three-dimensional super-resolution techniques, three-dimensional imaging will also be realized.

METHODS SUMMARY

Details of the experiment can be found in Methods. There we describe the fabrication of the aluminium film on a quartz slide, the self-assembly of silver nanoparticle clusters on a glass cover slide, the prism-TIRF experiment for the observation of the

hotspots on the aluminium film, the objective-TIRF experiment for the observation of the hotspots in silver nanoparticle clusters, the single-molecule localization method used to determine the adsorption centre and the fluorescence intensity of each single-molecule events, and the Gaussian kernel method for image rendering. More discussion on the single-molecule localization and the imaging rendering methods can be found in the Supplementary Information.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 28 June; accepted 18 November 2010.

1. Kneipp, K. *et al.* Single molecule detection using surface-enhanced Raman scattering (SERS). *Phys. Rev. Lett.* **78,** 1667–1670 (1997).
2. Nie, S. & Emory, S. R. Probing single molecules and single nanoparticles by surface-enhanced Raman scattering. *Science* **275,** 1102–1106 (1997).
3. Shalaev, V. M. & Stockman, M. I. Fractals: optical susceptibility and giant Raman scattering. *Z. Phys. D* **10,** 71–79 (1988).
4. Stockman, M. I., Faleev, S. V. & Bergman, D. J. Localization versus delocalization of surface plasmons in nanosystems: can one state have both characteristics? *Phys. Rev. Lett.* **87,** 167401 (2001).
5. Betzig, E. & Trautman, J. K. Near-field optics: microscopy, spectroscopy, and surface modification beyond the diffraction limit. *Science* **257,** 189–195 (1992).
6. Nelayah, J. *et al.* Mapping surface plasmons on a single metallic nanoparticle. *Nature Phys.* **3,** 348–353 (2007).
7. Vesseur, E. J. R., de Waele, R., Kuttge, M. & Polman, A. Direct observation of plasmonic modes in Au nanowires using high-resolution cathodoluminescence spectroscopy. *Nano Lett.* **7,** 2843–2846 (2007).
8. Kubo, A., Jung, Y. S., Kim, H. K. & Petek, H. Femtosecond microscopy of localized and propagating surface plasmons in silver gratings. *J. Phys. At. Mol. Opt. Phys.* **40,** S259–S272 (2007).
9. Jeanmaire, D. L. & Van Duyne, R. P. Surface Raman spectroelectrochemistry: Part I. Heterocyclic, aromatic, and aliphatic amines adsorbed on the anodized silver electrode. *J. Electroanal. Chem.* **84,** 1–20 (1977).
10. Albrecht, M. G. & Creighton, J. A. Anomalously intense Raman spectra of pyridine at a silver electrode. *J. Am. Chem. Soc.* **99,** 5215–5217 (1977).
11. Fleischmann, M., Hendra, P. J. & McQuillan, A. J. Raman spectra of pyridine adsorbed at a silver electrode. *Chem. Phys. Lett.* **26,** 163–166 (1974).
12. Creighton, J. A., Blatchford, C. G. & Albrecht, M. G. Plasma resonance enhancement of Raman scattering by pyridine adsorbed on silver or gold sol particles of size comparable to the excitation wavelength. *J. Chem. Soc. Faraday Trans. 2* **75,** 790–798 (1979).
13. Moskovits, M. Surface roughness and the enhanced intensity of Raman scattering by molecules adsorbed on metals. *J. Chem. Phys.* **69,** 4159–4161 (1978).
14. Gersten, J. & Nitzan, A. Electromagnetic theory of enhanced Raman scattering by molecules adsorbed on rough surfaces. *J. Chem. Phys.* **73,** 3023–3037 (1980).
15. Sarychev, A. K., Shubin, V. A. & Shalaev, V. M. Anderson localization of surface plasmons and nonlinear optics of metal-dielectric composites. *Phys. Rev. B* **60,** 16389–16408 (1999).
16. Seal, K. *et al.* Coexistence of localized and delocalized surface plasmon modes in percolating metal films. *Phys. Rev. Lett.* **97,** 206103 (2006).
17. Hutchison, J. A. *et al.* Subdiffraction limited, remote excitation of surface enhanced Raman scattering. *Nano Lett.* **9,** 995–1001 (2009).
18. Betzig, E. *et al.* Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313,** 1642–1645 (2006).
19. Rust, M. J., Bates, M. & Zhuang, X. W. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature Methods* **3,** 793–796 (2006).
20. Sharonov, A. & Hochstrasser, R. M. Wide-field subdiffraction imaging by accumulated binding of diffusing probes. *Proc. Natl Acad. Sci. USA* **103,** 18911–18916 (2006).
21. Wu, D., Liu, Z., Sun, C. & Zhang, X. Super-resolution imaging by random adsorbed molecule probes. *Nano Lett.* **8,** 1159–1162 (2008).
22. Hell, S. W. in *Single Molecule Spectroscopy in Chemistry, Physics and Biology* (eds Gräslund, A., Rigler, R. & Widengren, J.) 365–398 (Springer Series in Chemical Physics, Springer, 2009).
23. Roeffaers, M. *et al.* Super-resolution reactivity mapping of nanostructured catalyst particles. *Angew. Chem.* **121,** 9449–9453 (2009).
24. Mortensen, K. I., Churchman, L. S., Spudich, J. A. & Flyvbjerg, H. Optimized localization analysis for single-molecule tracking and super-resolution microscopy. *Nature Methods* **7,** 377–381 (2010).
25. Chang, C. C. *et al.* Aluminum oxidation in water. *J. Electrochem. Soc.* **125,** 787–792 (1978).
26. Wokaun, A., Lutz, H. P., King, A. P., Wild, U. P. & Ernst, R. R. Energy transfer in surface enhanced luminescence. *J. Chem. Phys.* **79,** 509–514 (1983).
27. Chance, R. R., Prock, A. & Silbey, R. Comments on the classical theory of energy transfer. *J. Chem. Phys.* **62,** 2245–2253 (1975).
28. Anger, P., Bharadwaj, P. & Novotny, L. Enhancement and quenching of single-molecule fluorescence. *Phys. Rev. Lett.* **96,** 113002 (2006).
29. Taminiau, T. H., Stefani, F. D., Segerink, F. B. & van Hulst, N. F. Optical antennas direct single-molecule emission. *Nature Photon.* **2,** 234–237 (2008).
30. Kinkhabwala, A. *et al.* Large single-molecule fluorescence enhancements produced by a bowtie nanoantenna. *Nature Photon.* **3,** 654–657 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank G. Bartal and A. Niv for discussions. This research was supported by the US Department of Energy Office of Science, Basic Energy Sciences and Lawrence Berkeley National Laboratory under contract no. DE-AC02-05CH11231.

Author Contributions H.C., A.L., X.Y. and X.Z. designed the experiments; H.C., A.L., C.G. and M.L. conducted experiments; C.L. and Y.L. conducted computer simulations and theoretical analysis; H.C., A.L., X.Y. and X.Z. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to X.Z. (xiang@berkeley.edu).

METHODS

Observations of hotspots on aluminium films. A quartz slide (3 inch × 1 inch, Ted Pella or SPI) is thoroughly cleaned according to a protocol developed at Radio Corporation of America (RCA protocol), before deposition. First, the slide is soaked in a bath of a base solution (distilled water, hydrogen peroxide and ammonium hydroxide at a ratio of 4:1:1) for 15 min; it is then transferred to an acid bath (distilled water, hydrogen peroxide and hydrogen chloride at a ratio of 6:1:1) for another 15 min, followed by 15 min of sonication in filtered distilled water (Milipore). The slide is then blow-dried with pure nitrogen gas and placed immediately inside an electron-beam evaporation chamber (Torr International). After the vacuum is pumped down to a minimum of 10^{-6} torr, deposition is started at a deposition rate of 0.1 nm s^{-1}. A crystal detector monitors the thickness of the film in real time. The deposition stops once the thickness reaches 15 nm.

A flow chamber is made by sandwiching an aluminium coated quartz slide and a clean glass cover slip using double-sided tape. The glass cover slips (no. 0, Ted Pella) are cleaned following the same procedure used for cleaning the quartz slide and stored in a clean jar filled with filtered distilled water before use. The chamber is placed on a home-built prism type total internal reflection (TIRF) microscope. A 200-mW 532-nm laser (Lambda Service) is used as the excitation light. The intensity is adjusted by neutral density filters to about 7 mW measured immediately before the prism. A 1-nm laser line band pass filter (Edmunds Optics), centred at 532 nm, is used to clean the excitation light that then impinges on a quartz Pellin-Broca prism (Thorlabs) and excites an evanescent wave on the surface of the sample. A 60×, NA 1.25 objective lens (Zeiss) is used to collect the light, and an electron multiplying charge coupled device (EMCCD Cascade 512, Photometric) is used to record fluorescent images. A 532-nm long pass filter (Semrock) and a band pass filter whose centre wavelength and bandwidth are 582 nm and 80 nm, respectively (582/80m from Chroma Technology), are placed after the objective lens. A 4× magnifier (Nikon) is placed right in front of the EMCCD in order to make the pixel size 72 nm. The camera's exposure time is 80 ms.

Filtered distilled water (Milipore) is first poured into the chamber. 10 nM Chromeo 542 dye (Active Motif) in filtered distilled water (Milipore) is then flowed into the chamber. Other dyes, including Cy3 (GE Healthcare), and Alexa-555 (Life technology) yield similar results. Immediately after the dye is flowed in, the fluorescence background increases significantly. Blinking at some spots was immediately observed and recorded. To avoid errors caused by the slow drifting of the objective lens, the total recording time is limited to a few minutes. To compare the enhancement factor, a similar experiment is performed in a quartz flow chamber without aluminium film coating, with a dye concentration of ~1 nM.

Observations of hotspots on silver nanoparticle clusters. Glass cover slips (no. 1.5, Fisher Scientific) are cleaned following the same procedure used for cleaning the quartz slide. The cleaned cover glass is stored in a clean jar filled with filtered distilled water before use. A drop of 0.1 ml concentrated solution of 40 nm silver colloids (Ted Pella) is coated on a cleaned cover glass. As the water evaporates, a layer of nanoparticle clusters formed on the surface. A home-built objective type TIRF with a Nikon TIRF objective (NA 1.49, 100×) is used. The excitation beam from a 40-mW 644-nm laser (Coherent, Cube) is reflected by a dichroic mirror with transition edge at 650 nm (650DCLP Chroma) into the objective lens. The power at the objective lens is measured to be ~4 mW. An emission filter (710/130m Chroma), whose centre and bandwidth are 710 nm and 130 nm respectively, is used before an EMCCD camera (Cascade 512, Princeton Instruments). To reduce the pixel size, a 4× magnifier (Nikon) is placed right in front of the EMCCD. The final pixel size corresponds to 48 nm per pixel. 10 nM Chromeo 642 dye (Active Motif) solution is used for the single molecule super-resolution imaging. The exposure time of the camera is 20 ms.

Single-molecule localization. This is performed by using a maximum likelihood estimation method[24]. Briefly, the log likelihood function

$$\sum_i (-E_i + n_i \log E_i - \log(n_i!))$$

where n_i is the observed number of photons at pixel i, and E_i is the expected number of photons calculated from the fitting parameters, is minimized. The accuracy of the estimation, determined from the variance, is:

$$\mathrm{Var}(x) = 2\frac{\sigma_a^2}{N}\left(1 + \int_0^1 \mathrm{d}t \frac{\log(t)}{1+t/\tau}\right)$$

Since each pixel of a camera has a finite size of a, the width of the Gaussian PSF has to be corrected as ${\sigma_a}^2 = \sigma^2 + a^2/12$. N is the total number of photons observed from the raw image. The integration in the bracket accounts for the shot noise from the background photons as $\tau = 2\pi{\sigma_a}^2 b^2/Na^2$, where b is the average background photon count[24]. The factor of *2* corrects the gain noise from an EMCCD camera.

We performed the single molecule localization in two steps: a least squares fitting of the raw image to a 2D Gaussian function is first used to obtain the initial parameters, and then the parameters are fed into the maximum likelihood estimation method to extract the centroid position x and y, and the amplitude, which are then used to generate the 3D scatter plot shown in Fig. 1.

Gaussian kernel image rendering. The collected single molecule events distribute in a random and sparse manner in space. As it is difficult to apply a fixed mesh directly for image rendering, we use a Gaussian Kernel rendering procedure.

Each image spot is analysed by a maximum likelihood estimation method[24] to extract the centroid position x, y and the intensity I. We then assign each single molecule event, represented by a colour sphere in the 3D scatter plot, to a Gaussian function, with the peak position as x and y, and the width as σ, determined from the accuracy of the single molecule localization by equation (1). The Gaussian function thus represents the probability of finding the true location of the single molecule event in space.

$$I(\mathbf{x}) = \sum I_i \exp\left(-(\mathbf{x}-\mathbf{x}_i)^2/2\sigma^2\right) / \sum \exp\left(-(\mathbf{x}-\mathbf{x}_i)^2/2\sigma^2\right)$$

Multiplying each Gaussian function with the intensity I_i from the fitting, summing them up, and normalizing the sum by the mean number of events, provides a smooth 3D image of the enhancement profile.

LETTER

doi:10.1038/nature09718

Grains and grain boundaries in single-layer graphene atomic patchwork quilts

Pinshane Y. Huang[1]*, Carlos S. Ruiz-Vargas[1]*, Arend M. van der Zande[2]*, William S. Whitney[2], Mark P. Levendorf[3], Joshua W. Kevek[4], Shivank Garg[3], Jonathan S. Alden[1], Caleb J. Hustedt[5], Ye Zhu[1], Jiwoong Park[3,6], Paul L. McEuen[2,6] & David A. Muller[1,6]

The properties of polycrystalline materials are often dominated by the size of their grains and by the atomic structure of their grain boundaries. These effects should be especially pronounced in two-dimensional materials, where even a line defect can divide and disrupt a crystal. These issues take on practical significance in graphene, which is a hexagonal, two-dimensional crystal of carbon atoms. Single-atom-thick graphene sheets can now be produced by chemical vapour deposition[1–3] on scales of up to metres[4], making their polycrystallinity almost unavoidable. Theoretically, graphene grain boundaries are predicted to have distinct electronic[5–8], magnetic[9], chemical[10] and mechanical[11–13] properties that strongly depend on their atomic arrangement. Yet because of the five-order-of-magnitude size difference between grains and the atoms at grain boundaries, few experiments have fully explored the graphene grain structure. Here we use a combination of old and new transmission electron microscopy techniques to bridge these length scales. Using atomic-resolution imaging, we determine the location and identity of every atom at a grain boundary and find that different grains stitch together predominantly through pentagon–heptagon pairs. Rather than individually imaging the several billion atoms in each grain, we use diffraction-filtered imaging[14] to rapidly map the location, orientation and shape of several hundred grains and boundaries, where only a handful have been previously reported[15–19]. The resulting images reveal an unexpectedly small and intricate patchwork of grains connected by tilt boundaries. By correlating grain imaging with scanning probe and transport measurements, we show that these grain boundaries severely weaken the mechanical strength of graphene membranes but do not as drastically alter their electrical properties. These techniques open a new window for studies on the structure, properties and control of grains and grain boundaries in graphene and other two-dimensional materials.

Figure 1a shows a large array of the suspended, single-layer graphene membranes used in this study. We grew predominately single-layer graphene films on copper foils by chemical vapour deposition[1] (CVD) using three different growth recipes, which we refer to as growth methods A, B and C. Unless otherwise stated, all data were taken on graphene grown with method A, which was similar to the recipe reported in ref. 1. Methods B and C are slight variations: method B uses ultrapure copper foils[18] (99.999% pure rather than 99.8%) and method C uses a rapid thermal processor furnace (Methods). These films were transferred onto holey silicon nitride or Quantifoil transmission electron microscopy (TEM) grids using two different techniques (Methods and Supplementary Information). One key innovation over previous graphene TEM sample fabrication[20] was the gentle transfer of the graphene onto a TEM grid using a minimum of polymer support and baking the samples in air to remove the polymer without liquid solvents. This produces large arrays of free-standing graphene sheets covering up to 90% of the TEM grid holes.

To characterize these membranes at the atomic scale, we used aberration-corrected annular dark-field scanning transmission electron microscopy (ADF-STEM), where a 60-keV, ångström-scale electron beam is scanned over the sample while the medium- to high-angle scattered electrons are collected. Keeping the electron beam voltage below the ~100-keV graphene damage threshold was necessary to limit beam-induced damage. Properly calibrated, this technique images the location and atomic number[21] of each atom and, along with TEM, has been used to study the lattice and atomic defects of graphene and boron nitrene[19,21–23]. Figure 1b shows an ADF-STEM image of the crystal lattice within a single graphene grain. Away from the grain boundaries, such regions are defect free.

In Fig. 1c, two graphene grains meet with a relative misorientation of 27°, forming a tilt boundary. Additional images of grain boundaries are shown in Supplementary Figs 2c and 3. As highlighted in Fig. 1d, the two crystals are stitched together by a series of pentagons, heptagons and distorted hexagons. The grain boundary is not straight, and the defects along the boundary are not periodic. Although the boundary dislocation

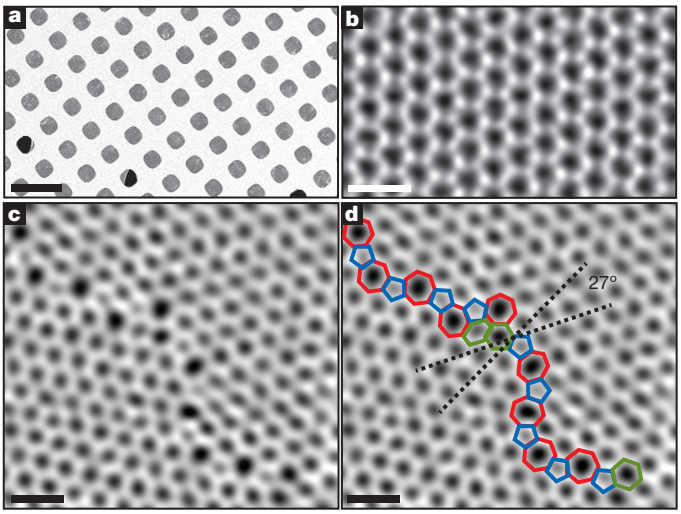


Figure 1 | Atomic-resolution ADF-STEM images of graphene crystals. **a**, Scanning electron microscope image of graphene transferred onto a TEM grid with over 90% coverage using novel, high-yield methods. Scale bar, 5 μm. **b**, ADF-STEM image showing the defect-free hexagonal lattice inside a graphene grain. **c**, Two grains (bottom left, top right) intersect with a 27° relative rotation. An aperiodic line of defects stitches the two grains together. **d**, The image from **c** with the pentagons (blue), heptagons (red) and distorted hexagons (green) of the grain boundary outlined. **b**–**d** were low-pass-filtered to remove noise; scale bars, 5 Å.

[1]School of Applied and Engineering Physics, Cornell University, Ithaca, New York 14853, USA. [2]Department of Physics, Cornell University, Ithaca, New York 14853, USA. [3]Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, USA. [4]Department of Physics, Oregon State University, Corvallis, Oregon 97331, USA. [5]Department of Physics and Astronomy, Brigham Young University, Provo, Utah 84602, USA. [6]Kavli Institute at Cornell for Nanoscale Science, Ithaca, New York 14853, USA.
*These authors contributed equally to this work.

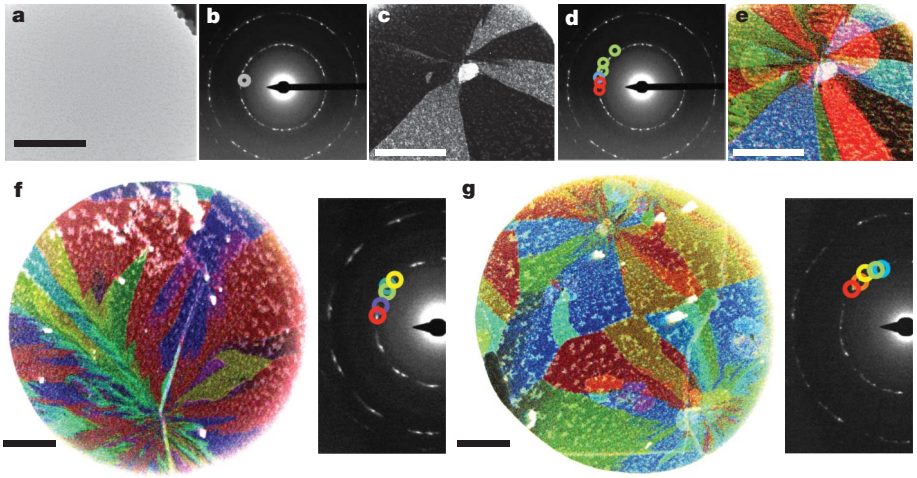


Figure 2 | Large-scale grain imaging using DF-TEM. **a**–**e**, Grain imaging process. **a**, Samples appear uniform in bright-field TEM images. **b**, Diffraction pattern taken from a region in **a** reveals that this area is polycrystalline. Placing an aperture in the diffraction plane filters the scattered electrons forming **c**, a corresponding dark-field image showing the real-space shape of these grains. **d**, Using several different aperture locations and colour-coding them produces **e**, a false-colour, dark-field image overlay depicting the shapes and lattice orientations of several grains. **f**, **g**, Images of regions where many grains emanate from a few points. Scale bars, 500 nm.

resembles structures proposed theoretically[11,13], its aperiodicity contrasts with many of these models and will strongly affect the predicted properties of grain boundaries. By analysing atomic scattering intensities[21], we confirm that the boundary is composed entirely of carbon. In addition, although high electron beam doses could induce isolated bond rotations (Supplementary Fig. 3), the boundary was largely stable under the 60-keV electron beam. Thus, the polycrystalline graphene is a strongly bonded, continuous carbon membrane. We also note that many grain boundaries are decorated by lines of surface particles and adsorbates (Supplementary Fig. 4), suggesting that, as predicted[10], they may be more chemically reactive than the pristine graphene lattice.

Both STEM and TEM, which determine the positions and identities of atomic nuclei, and complementary scanning tunnelling microscopy, used to probe valence wavefunctions[15–17], are invaluable for understanding the local properties of grain boundaries. Using these atomic-resolution approaches, however, tens of billions to hundreds of billions of pixels would be needed to image even a single micrometre-scale grain fully, with estimated acquisition times of a day or more. Other candidates for characterizing grains on larger scales, such as low-energy electron microscopy[18] and Raman microscopy[3], typically cannot resolve small grains and may be difficult to interpret. Fortunately, electron microscopy offers an ideal technique for imaging grains on the necessary length scales: dark-field TEM (DF-TEM), which is a high-throughput, diffraction-sensitive imaging technique[14] that can be implemented on most TEMs built in the past sixty years. This method is usually applied to foils about 100–300-nm thick[14], but we demonstrate below that, remarkably, it also works on single-atom-thick sheets—even on samples too dirty for atomic-resolution imaging. In this manner, DF-TEM provides a nanometre- to micrometre-scale grain analysis that complements ADF-STEM to give a complete understanding of graphene grains on every relevant length scale.

Figure 2a, b shows a bright-field TEM image of a graphene sheet along with the selected-area electron diffraction pattern created from this region of the membrane. Owing to graphene's six-fold symmetry, electron diffraction from a single graphene crystal results in one set of six-fold-symmetric spots. Figure 2b contains many such families of spots, indicating that the field of view contains several grains of different orientations. DF-TEM images these grains one by one with few-nanometre resolution using an objective aperture filter in the back focal plane to collect electrons diffracted through a small range of angles, as shown by the circle in Fig. 2b. The resulting real-space image (Fig. 2c) shows only the grains corresponding to these selected in-plane lattice orientations and requires only a few seconds to acquire. By repeating this process using several different aperture filters, then colouring and overlaying these dark-field images (Fig. 2d, e), we create complete maps of the graphene grain structure, colour-coded by lattice orientation, as shown in Fig. 2e–g.

The images obtained are striking. The grains have complex shapes and many different crystal orientations. In Fig. 2e–g, we observe special locations from which many grains emanate. Small particles and multilayer graphene also are often found near these sites; see, for example, Fig. 2e, top right. Both the average spacing (2–4 μm) and shapes of these radiant sites when we use growth method A are comparable with Raman and scanning electron microscope observations of graphene nucleation[1,3], suggesting that these locations are probably nucleation sites. Similar structures have been observed in studies of crystallization in colloids and are consistent with crystallization around impurities[24]. Similar multigrain nucleation on copper has recently been observed using low-energy electron microscopy[18]. Significantly, each apparent nucleation site gives rise to many grains of different orientations, resulting in a mean grain size much smaller than the nucleation density.

The distributions of grain size and relative angular orientation are readily determined from DF-TEM images. As discussed below, grain sizes are dependent on growth conditions, here ranging from hundreds of nanometres to tens of micrometres for slight changes in growth conditions. In Fig. 3a, we plot a histogram of grain sizes across several samples grown using method A. The mean grain size, defined as the square root of the grain area, is 250 ± 11 nm (s.e.m.). This size is much smaller than the grain size of the copper substrate[1,4] (100 μm–1 mm) and typical lateral grains measured in bulk, highly ordered pyrolytic graphite[25] (6–30 μm). The inset in Fig. 3a shows the cumulative probability of finding multiple grains in a given area. This plot demonstrates that micrometre-scale CVD graphene devices produced from this set of films will nearly always contain multiple grains. Figure 3b shows a histogram of the relative crystallographic angles between adjacent grains. Because of graphene's six-fold crystal symmetry, the diffractive imaging technique only determines grain rotations modulo 60°. Consequently, the measurable difference between grain orientations is from 0 to 30° (with, for example, 31° measured as 29°). We observe a surprising and robust preference for low-angle (~7°) grain boundaries and high-angle (~30°) boundaries similar to that seen in Fig. 1.

Additional information about these orientations comes from the larger-area diffraction patterns in Fig. 3c, which we created by averaging diffraction data sampled across 1,200-μm^2 regions of graphene. The broadened diffraction peaks in Fig. 3c (left) show a distinct six-fold pattern, indicating that a significant fraction of the grains are approximately aligned across large areas This alignment can also be seen in Fig. 3d, which is a low-magnification DF-TEM image showing grains with a small (~10°) range of in-plane lattice orientations. Almost half of the membrane appears bright, indicating that these grains are all approximately aligned. In contrast, a dark-field image of randomly oriented grains would only show roughly one-sixth (10°/60°) of the graphene membrane. In the diffraction pattern of a separately grown sample (Fig. 3c, right), we instead find a clear 12-fold periodicity,

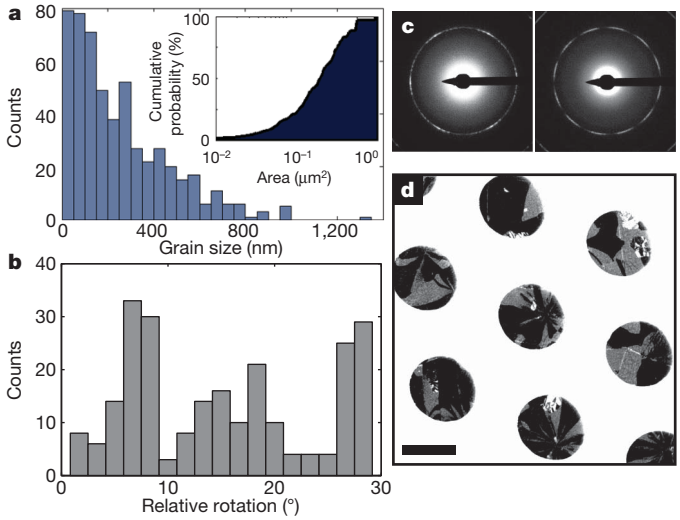


Figure 3 | Statistical analysis of grain size and orientation. **a**, Histogram of grain sizes, taken from three representative samples using DF-TEM. The mean grain size is 250 ± 11 nm (s.e.m., $n = 535$). Inset, plot of the cumulative probability of having more than one grain given the area of a device. **b**, Histogram of relative grain rotation angles measured from 238 grain boundaries. **c**, **d**, Large-area diffraction patterns (**c**) and a low-magnification DF-TEM image (**d**) show that grains are globally aligned near particular directions. Scale bar, 2 μm.

indicating that there are two main families of grains rotated from one another by 30°. These distributions, which often contain smaller subpeaks (Supplementary Fig. 6), are consistent with the frequent observation of low-angle and high-angle (~30°) grain boundaries. We attribute these alignments to registry to the copper substrate used for graphene growth. Such registry has recently been observed in low-energy electron microscopy and scanning tunnelling microscopy studies of graphene growth on copper (100) and (111) surfaces[15,16,18].

By directly correlating grain structure with growth methods, these DF-TEM methods can be used to build on recent studies[3] that have demonstrated links between island nucleation density and growth conditions. Fig. 4a–c shows three composite DF-TEM images of graphene grown using methods A, B and C. The slight differences between growth methods effected significant changes in the grain size, shape and crystallographic orientation of the CVD graphene. For example, with growth method C we observed grains averaging 1–4 μm (Fig. 4c), which is an order of magnitude larger than the grains grown using method A. Our DF-TEM methods provide a powerful characterization tool for understanding and controlling grain growth, which will be a rich field of study important for graphene applications.

The ability to image the grain structure in graphene monolayers easily opens the door to the systematic exploration of the effects of grain structure on the physical, chemical, optical and electronic properties of graphene membranes. We find that such studies are further facilitated because grain boundaries are visible in scanning electron microscopy and atomic force microscopy (AFM) phase imaging owing to preferential decoration of the grain boundaries with surface contamination (Fig. 5a and Supplementary Figs 9 and 10). Below, we show two examples probing the electrical and mechanical properties of grain boundaries.

We first examine the failure strength of the polycrystalline CVD graphene membranes (growth method A) using AFM. We used AFM phase imaging to image grains (Fig. 5a) and then pressed downwards with the AFM tip to test the mechanical strength of the membranes. As seen in Fig. 5b, the graphene tears along the grain boundaries. From repeated measurements, we find that failure occurs at loads of ~100 nN, which is an order of magnitude lower than typical fracture loads of 1.7 μN reported for single-crystal exfoliated graphene[26]. Thus, the strength of polycrystalline graphene is dominated by its grain boundaries.

We probed the electrical properties of polycrystalline graphene by fabricating electrically contacted devices using graphene from all three growth methods. Figure 4d shows a histogram of mobilities extracted from four-point transport measurements. Devices grown using methods A, B and C have room-temperature mobilities of 1,000 ± 750, 7,300 ± 1,100 and 5,300 ± 2,300 $cm^2 V^{-1} s^{-1}$(s.d.), respectively. The mobilities for growth method A are comparable to previous results on CVD graphene[1], whereas the mobilities of growth methods B and C are closer to those reported for exfoliated graphene[27] (1,000–20,000 $cm^2 V^{-1} s^{-1}$). By comparing these measurements with the corresponding DF-TEM images in Fig. 4a–c, we are surprised to find that, although mobility is clearly affected by growth conditions, high mobility does not directly correlate with large grain size.

To complement these bulk electrical measurements, we used scanning probe a.c. electrostatic force microscopy[28] (AC-EFM) to test the resistivity of individual grain boundaries. We fabricated suspended graphene membrane devices[29]. One of these is shown schematically in Fig. 5c, where we also plot the relative potential along a graphene membrane between two biased electrodes, measured using AC-EFM. In this plot, high-resistance grain boundaries would manifest as sharp drops in potential. The graphene in these devices (growth method A) had a mean grain size of 250 nm, so a line scan across these 3-μm-long membranes should cross an average of 12 grains. However, no noticeable potential drops were detected, indicating that most grain boundaries in these devices are not strongly resistive interfaces. By assuming that the grain boundary runs perpendicular to the line scan, we estimate an upper bound on the grain boundary resistance of $R_{GB} < 60\ \Omega\ \mu m/L$, where L is the length of the grain boundary, to be compared with the sheet resistance of $R_{graphene} = 700\ \Omega/\square$ for the entire device. In other words, the resistance of the grain boundaries is less than one-third the resistance of a 250-nm grain. Further measurements on six additional graphene membranes, both suspended and unsuspended, and from different growth methods, produced similar results. This small impact of grain boundaries stands in stark contrast to other materials, such as complex oxides, where a single grain boundary can lead to a million-fold increase in resistance over single crystals[30].

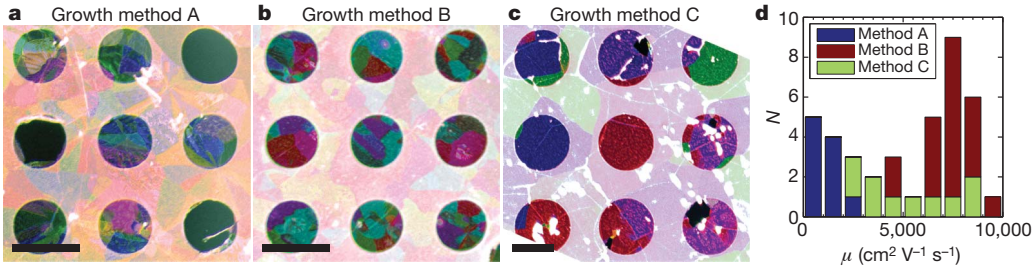


Figure 4 | Grain structure and mobilities for three growth conditions. **a–c**, Composite DF-TEM images of grain structure show variations with growth condition. The mean grain sizes are 250 ± 11 nm (s.e.m.; growth method A, 99.8% pure copper), 470 ± 36 nm (s.e.m.; growth method B, 99.999% pure (ultrapure) copper) and 1.7 ± 0.15 μm (s.e.m.; growth method C (rapid thermal anneal)). The graphene is visible through the 20-nm, perforated amorphous-carbon Quantifoil support film. The graphene is broken over three of the perforations in **a**. Scale bars, 2 μm. **d**, Vertically stacked histogram of room-temperature mobilities, μ, measured from 39 devices using graphene growth methods A, B, and C. N, number of devices. See Methods for further details.

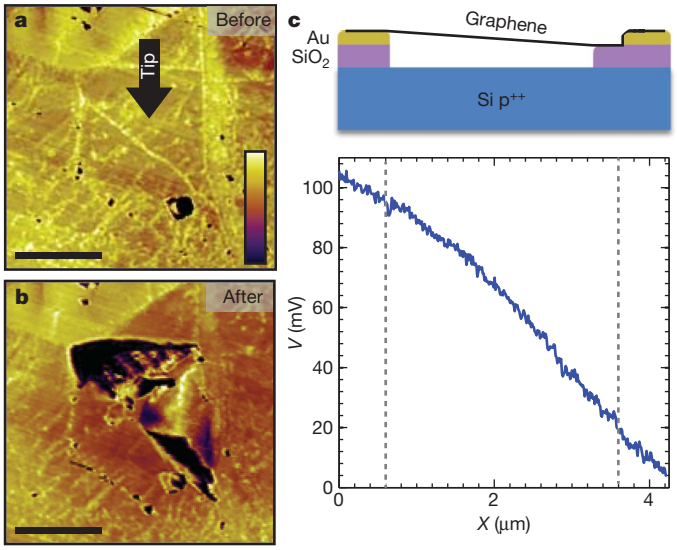


Figure 5 | AFM indentation and AC-EFM studies of graphene grain boundaries. **a**, **b**, AFM phase images of a graphene grain before and after an indentation measurement. **a**, Indentation takes place at the centre of this grain as shown by the arrow. **b**, The region is torn along grain boundaries after indentation. Scale bars, 200 nm. **c**, Electrostatic potential, averaged over three adjacent line scans along a suspended graphene sheet between two electrodes (schematic at top) and measured using AC-EFM. Although on average each line scan should cross 12 grains, no measureable features are present. Dashed lines indicate the locations of the electrodes.

The imaging techniques reported here provide the tools to characterize graphene grains and grain boundaries on all relevant length scales. These methods will be crucial both for exploring synthesis strategies to optimize grain properties and for studies, such as those described above, on the microscopic and macroscopic impact of grain structure on graphene membranes. Thus, these results represent a significant step forward in realizing the ultimate promise of atomic membranes in electronic, mechanical and energy-harvesting devices.

METHODS SUMMARY

TEM/STEM. We did ADF-STEM imaging using a NION UltraSTEM100 with imaging conditions similar to those used in ref. 21. At 60 kV, using a 33–35-mrad convergence angle, the electron probe was close to 1.3 Å in size and did not damage the graphene. Images presented in Figs 1–4 were acquired with the medium-angle annular dark-field detector with acquisition times of between 16 and 32 μs per pixel. For TEM imaging, we used a FEI Technai T12 operated at 80 kV. Acquisition times for dark-field images were 5–10 s per frame. The spatial resolution in dark-field images ranges from 1 to 10 nm and is set by the size of the objective filtering aperture, in a trade-off between real-space resolution and angular resolution in reciprocal space.

Scanning probe measurements. For AFM deflection measurements, we used a MFP3D scope from Asylum Research. We used silicon AFM probes (Multi75Al, Budget Sensors) with a resonant frequency of ~75 kHz, a force constant of ~3 N m^{-1} and a tip radius of <10 nm. All imaging was done in tapping mode. For AC-EFM measurements, a DI 4100 AFM with a signal access module was operated in lift mode with a constant probe tip voltage, $V_{tip} = 2$ V, a lift height of 10 nm and no piezo drive on the tip. An a.c. voltage of $V_0 = 1$ V was applied through the electrodes at the resonance frequency of the EFM cantilever, $f_{cant} \approx 77$ kHz. An electrostatic force drives the EFM cantilever to resonate, and the amplitude of motion is measured.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 23 September; accepted 29 November 2010.
Published online 5 January 2011.

1. Li, X. *et al.* Large-area synthesis of high-quality and uniform graphene films on copper foils. *Science* **324,** 1312–1314 (2009).
2. Reina, A. *et al.* Large area, few-layer graphene films on arbitrary substrates by chemical vapor deposition. *Nano Lett.* **9,** 30–35 (2009).
3. Li, X. *et al.* Graphene films with large domain size by a two-step chemical vapor deposition process. *Nano Lett.* **10,** 4328–4334 (2010).
4. Bae, S. *et al.* Roll-to-roll production of 30-inch graphene films for transparent electrodes. *Nature Nanotechnol.* **5,** 574–578 (2010).
5. Cervenka, J. & Flipse, C. F. J. Structural and electronic properties of grain boundaries in graphite: planes of periodically distributed point defects. *Phys. Rev. B* **79,** 195429 (2009).
6. Peres, N. M. R., Guinea, F. & Castro-Neto, A. H. Electronic properties of disordered two-dimensional carbon. *Phys. Rev. B* **73,** 125411 (2006).
7. Yazyev, O. V. & Louie, S. G. Electronic transport in polycrystalline graphene. *Nature Mater.* **6,** 806–809 (2010).
8. Mesaros, A., Papanikolaou, S., Flipse, C. F. J., Sadri, D. & Zaanen, J. Electronic states of graphene grain boundaries. *Phys. Rev. B* **82,** 205119 (2010).
9. Cervenka, J., Katsnelson, M. I. & Flipse, C. F. J. Room-temperature ferromagnetism in graphite driven by two-dimensional networks of point defects. *Nature Phys.* **5,** 840–844 (2009).
10. Malola, S., Hakkinen, H. & Koskinen, P. Structural, chemical, and dynamical trends in graphene grain boundaries. *Phys. Rev. B* **81,** 165447 (2010).
11. Liu, Y. & Yakobson, B. I. Cones, pringles, and grain boundary landscapes in graphene topology. *Nano Lett.* **10,** 2178–2183 (2010).
12. Grantab, R., Shenoy, V. B. & Ruoff, R. S. Anomalous strength characteristics of tilt grain boundaries in graphene. *Science* **330,** 946–948 (2010).
13. Yazyev, O. V. & Louie, S. G. Topological defects in graphene: dislocations and grain boundaries. *Phys. Rev. B* **81,** 195420 (2010).
14. Hirsch, P., Howie, A., Nicholson, R., Pashley, D. W. & Whelan, M. J. *Electron Microscopy of Thin Crystals* (Krieger, 1965).
15. Zhao, L. *et al.* The atomic-scale growth of large-area monolayer graphene on single-crystal copper substrates. Preprint at ⟨http://arxiv.org/abs/1008.3542⟩ (2010).
16. Gao, L., Guest, J. R. & Guisinger, N. P. Epitaxial graphene on Cu(111). *Nano Lett.* **10,** 3512–3516 (2010).
17. Cockayne, E. *et al.* Rotational grain boundaries in graphene. Preprint at ⟨http://arxiv.org/abs/1008.3574⟩ (2010).
18. Wofford, J. M., Nie, S., McCarty, K. F., Bartlet, N. C. & Dubon, O. D. Graphene islands on Cu foils: the interplay between shape, orientation, and defects. *Nano Lett.* **10,** 4890–4896 (2010).
19. Park, H. J., Meyer, J., Roth, S. & Skakalova, V. Growth and properties of few-layer graphene prepared by chemical vapor deposition. *Carbon* **48,** 1088–1094 (2010).
20. Regan, W. *et al.* A direct transfer of layer-area graphene. *Appl. Phys. Lett.* **96,** 113102 (2010).
21. Krivanek, O. L. *et al.* Atom-by-atom structural and chemical analysis by annular dark-field electron microscopy. *Nature* **464,** 571–574 (2010).
22. Hashimoto, A., Suenaga, K., Gloter, A., Urita, K. & Iijima, S. Direct evidence for atomic defects in graphene layers. *Nature* **430,** 870–873 (2004).
23. Meyer, J. C. *et al.* Direct imaging of lattice atoms and topological defects in graphene membranes. *Nano Lett.* **8,** 3582–3586 (2008).
24. de Villeneuve, V. W. A. *et al.* Hard sphere crystal nucleation and growth near large spherical impurities. *J. Phys. Condens. Matter* **17,** S3371–S3378 (2005).
25. Park, S., Floresca, H. C., Suh, Y. & Kim, M. J. Electron microscopy analyses of natural and highly oriented pyrolytic graphites and the mechanically exfoliated graphenes produced from them. *Carbon* **48,** 797–804 (2010).
26. Lee, C., Wei, X., Kysar, J. W. & Hone, J. Measurement of the elastic properties and intrinsic strength of monolayer graphene. *Science* **321,** 385–388 (2008).
27. Geim, A. K. & Novoselov, K. S. The rise of graphene. *Nature Mater.* **6,** 183–191 (2007).
28. Bachtold, A. *et al.* Scanned probe microscopy of electronic transport in carbon nanotubes. *Phys. Rev. Lett.* **84,** 6082–6085 (2000).
29. Van Der Zande, A. M. *et al.* Large-scale arrays of single-layer graphene resonators. *Nano Lett.* **10,** 4869–4873 (2010).
30. Thiel, S. *et al.* Electron scattering at dislocations in LaAlO3/SrTiO3 interfaces. *Phys. Rev. Lett.* **102,** 046809 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors acknowledge discussions with M. Blees, J. Cha, S. Gerbode, J. Grazul, E. Kirkland, L. Fitting-Kourkoutis, O. Krivanek, S. Shi, S. Wang and H. Zhuang. This work was supported in part by the National Science Foundation through the Cornell Center for Materials Research and the Nanoscale Science and Engineering Initiative of the National Science Foundation under NSF Award EEC-0117770, 064654. Additional support was provided by the Army Research Office, CONACYT-Mexico, the Air Force Office of Scientific Research, DARPA-MTO and the MARCO Focused Research Center on Materials, Structures, and Devices. Sample fabrication was performed at the Cornell node of the National Nanofabrication Infrastructure Network, funded by the NSF. Additional facilities support was provided by the Cornell Center for Materials Research (NSF DMR-0520404 and IMR-0417392) and NYSTAR.

Author Contributions P.Y.H., C.S.R.-V. and A.M.v.d.Z. contributed equally to this work. Electron microscopy and data analysis were carried out by P.Y.H. and D.A.M., with Y.Z. contributing to initial DF-TEM. Graphene growth and sample fabrication were done by A.M.v.d.Z. and C.S.R.-V. under the supervision of P.L.M. and J.P., aided by M.P.L., S.G., W.S.W., J.W.K., J.S.A. and C.J.H. AC-EFM, mobility measurements and analysis were done by A.M.v.d.Z. and P.L.M., aided by C.S.R.-V. and J.W.K. AFM mechanical testing and analysis were done by C.S.R.-V. and J.P., aided by S.G. All authors discussed the results and implications at all stages. P.Y.H, A.M.v.d.Z., C.S.R.-V., P.L.M., J.P. and D.A.M. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.A.M. (david.a.muller@cornell.edu).

METHODS

ADF-STEM. ADF-STEM imaging was conducted using a NION UltraSTEM100 operated at 60 kV. Imaging conditions were similar to those used in ref. 21. Using a 33–35-mrad convergence angle, our probe size was close to 1.3 Å. Because the low-voltage electron beam was below the damage threshold energy[31], the pristine graphene lattice remains stable and defect free. High electron beam doses could induce isolated bond rotations at grain boundaries (Supplementary Fig. 4) similar to those seen in ref. 32. Images presented in Figs 1–4 were acquired with the medium-angle annular dark-field detector with acquisition times of between 16 and 32 μs per pixel.

DF-TEM. TEM imaging was conducted using a FEI Technai T12 operated at 80 kV, which did not cause any apparent damage to the graphene membranes. Acquisition time for dark-field images were 5–10 s per frame. The spatial resolution in dark-field images ranges from 1 to 10 nm and is set by the size of the objective filtering aperture in a trade-off between real-space resolution and angular resolution in reciprocal space.

AC-EFM. A DI 4100 AFM with a signal access module was operated in lift mode with tip voltage $V_{tip} = 2$ V, a lift height of 10 nm and no piezo drive on the tip. An a.c. voltage $V_0 = 1$ V was applied through the electrodes at the resonance frequency of the EFM cantilever, $f_{cant} \approx 77$ kHz. An electrostatic force drives the EFM cantilever to resonate, and the amplitude of motion is measured.

AFM imaging and deflection measurements. For AFM deflection measurements, we used a MFP3D scope from Asylum Research. We used silicon AFM probes (Multi75Al, Budget Sensors) with a resonant frequency of ~75 kHz, a force constant of ~3 N m^{-1} and a tip radius of <10 nm. All imaging was done in tapping mode. Images were taken with resolutions of 512 × 512 or 1,024 × 1,024, with acquisition times of at most 10 min.

Graphene growth. We grew single-layer graphene using CVD on copper foils in three ways. Growth method A: similar to methods described in ref. 1, we annealed a 99.8% pure copper foil (Alfa Aesar #13382) at 1,000 °C at low pressure with an H_2 flow of 7 standard cubic centimetres per minute (s.c.c.m.) for 10 min. We then grew the graphene at 1,000 °C by flowing CH_4:H_2 at 150:7 s.c.c.m. for 10–15 min (varying growth time within this range did not yield noticeably different results). Samples are cooled for ~50 min while the CH_4:H_2 flow is maintained. Growth method B: this is identical to method A, except we used higher purity (99.999%) copper foil (Alfa Aesar #10950). Growth method C: we used a rapid thermal processor tube furnace with a ~4″ inner diameter (MTI Corporation). We annealed copper foil (99.8% purity) at 1,000 °C (H_2, 300 s.c.c.m.) for 30 min, and then grew the graphene at 1,000 °C (CH_4:H_2, 875:300 s.c.c.m.) for 60 min.

Samples for DF-TEM. We transferred the graphene either to commercial holey SiN TEM grids (such as PELCO Holey Silicon Nitride Support Films) with 2.5-μm-diameter holes or to Quantifoil holey carbon TEM grids to allow imaging of larger grains. Quantifoil grids are typically 10–20 nm thick, which is thin enough to allow DF-TEM imaging through the carbon support.

The fabrication for DF-TEM samples is a gentle graphene transfer method using a thin PMMA support, which produced roughly 90% coverage of TEM grid holes (that is, 90% of grid holes were uniformly covered with suspended graphene). After graphene growth on a copper foil, a thin layer of PMMA was spun onto the graphene (2% in anisole, 4,000 r.p.m. for 30 s), without a post-baking step. Copper was then etched away by floating the foil, PMMA side up, in a HCl/$FeCl_3$ copper etchant (Transene, Type 100/200). Next, the graphene and polymer support were washed by transferring them to deionized-water baths, taking care to not bring the PMMA into contact with liquids, to avoid depositing unwanted residues on the PMMA side of this layer. Finally, the PMMA–graphene layer is scooped out in pieces onto TEM grids. PMMA can be thermally decomposed[33], which is a gentler process than using liquid solvent rinses. We baked our samples in air (350 °C for 3–4 h), without the use of an argon flow, which can slow the cleaning effect substantially. This step removes the PMMA layer, leaving the graphene freely suspended in a liquid-free release process. These high-yield samples were used in DF-TEM because they provided enough clean graphene to image large numbers of grains.

Samples for ADF-STEM. Our secondary technique produced cleaner, but lower-yield, graphene using a polymer-free transfer method. This technique is similar to the methods of ref. 20, in which TEM grids are placed on top of the foil before etching and attached by dropping methanol on the grids. Our main addition to this technique was to bake the final samples in a series of annealing processes increasing in temperature. The grids were then baked in air at 350 °C for 2 h. In this method, the samples are annealed in ultrahigh vacuum by ramping the temperature to 950 °C, holding this temperature steady for 15 min and then cooling to room temperature without active cooling. This annealing is done below the graphene growth temperature, and the micrometre-scale grain structure did not change afterwards. Thus, any change that may result from annealing should be small in comparison with changes occurring during the formation of the grain boundaries. A final step was to anneal the grids at 130 °C for >8 h before transferring them in air to the TEM. Because this transfer method uses no support film for the graphene as it is transferred, this method was a comparatively low-yield transfer process with coverage of just a few per cent over the holes. The advantage to this technique over the polymer-based transfer is that it produced graphene with less surface carbon contamination—regions hundreds of nanometres wide appeared atomically clean in ADF-STEM images.

Electrically contacted samples. We fabricated top-gated graphene devices in four-point probe geometry (shown in Supplementary Fig. 11a, b, with electrodes labelled). A transferred graphene film was patterned by photolithography and a 10-s exposure to an oxygen plasma to define the graphene strips. This was followed by fabricating 1.5-nm Ti/4.5-nm Au electrodes. We patterned a top gate, to measure the charge mobility in graphene, by electron beam evaporation first of 90 nm of silicon oxide as a dielectric layer and then of a Cr/Au layer (1.5 nm/50 nm), without breaking vacuum between each evaporation.

For the EFM measurements, we fabricated electrically contacted, suspended graphene by growing single-layer graphene on copper using CVD; patterning the graphene into 3-μm-wide strips while still on the copper foil, using contact lithography; and transferring the patterned strips onto a substrate with pre-patterned gold electrodes and trenches.

31. Meyer, J. C., Chuvilin, A. & Kaiser, U. in *MC2009, Vol. 3: Materials Science* (eds Grogger, W., Hofer, F. & Polt, P.) 347–348 (Graz Univ. Technology, 2009).
32. Suenaga, K. *et al.* Imaging active topological defects in carbon nanotubes. *Nature Nanotechnol.* **2,** 358–360 (2007).
33. Jiao, L. *et al.* Creation of nanostructures with poly(methyl methacrylate)-mediated nanotransfer printing. *J. Am. Chem. Soc.* **130,** 12612–12613 (2008).

LETTER

doi:10.1038/nature09668

Primitive agriculture in a social amoeba

Debra A. Brock[1], Tracy E. Douglas[1], David C. Queller[1] & Joan E. Strassmann[1]

Agriculture has been a large part of the ecological success of humans[1]. A handful of animals, notably the fungus-growing ants, termites and ambrosia beetles[2–4], have advanced agriculture that involves dispersal and seeding of food propagules, cultivation of the crop and sustainable harvesting[5]. More primitive examples, which could be called husbandry because they involve fewer adaptations, include marine snails farming intertidal fungi[6] and damselfish farming algae[7]. Recent work has shown that microorganisms are surprisingly like animals in having sophisticated behaviours such as cooperation, communication[8,9] and recognition[10,11], as well as many kinds of symbiosis[12–15]. Here we show that the social amoeba *Dictyostelium discoideum* has a primitive farming symbiosis that includes dispersal and prudent harvesting of the crop. About one-third of wild-collected clones engage in husbandry of bacteria. Instead of consuming all bacteria in their patch, they stop feeding early and incorporate bacteria into their fruiting bodies. They then carry bacteria during spore dispersal and can seed a new food crop, which is a major advantage if edible bacteria are lacking at the new site. However, if they arrive at sites already containing appropriate bacteria, the costs of early feeding cessation are not compensated for, which may account for the dichotomous nature of this farming symbiosis. The striking convergent evolution between bacterial husbandry in social amoebas and fungus farming in social insects makes sense because multigenerational benefits of farming go to already established kin groups.

The social amoeba *Dictyostelium discoideum* is well known for its social interactions. When prey bacteria become scarce, amoebae aggregate by the tens of thousands and produce a multicellular migratory slug that becomes a fruiting body in which about 20% of cells die to form a sterile stalk. The stalk aids the dispersal of the remaining cells, which differentiate into spores in a spherical structure called the sorus[16,17] (Fig. 1a). We show that about one-third of wild-collected clones husband bacteria through the sporulation and dispersal process. We call these clones farmers because they carry, seed and prudently harvest their food, but the farming is primitive because no active cultivation is known.

At first glance, beneficial interactions occurring between *D. discoideum* and the bacteria would be unexpected. In the unicellular state, *D. discoideum* are solitary predators of bacteria[18], and bacteria use a wide range of strategies to deter predation[19]. Early life-history work on another species, *Dictyostelium mucoroides*, suggested the possibility of symbiosis with bacteria[20,21] but did not document this interaction experimentally. The possibility fell out of favour, particularly after work[18] finding no support for a symbiotic relationship and much support for a predatory one. However, that study considered only a single *D. discoideum* clone, and the extent of natural variation remains largely unexplored. In our study, we used a population of 35 wild *D. discoideum* clones isolated from soil collected at Mountain Lake Biological Station, Virginia, and Lake Itasca Biological Station, Minnesota (Supplementary Table 1). We observed that the sorus contents of some of the clones contained bacteria in addition to *D. discoideum* spores (Fig. 1b). To confirm this observation, we initially chose four clones that seemed to have bacteria in their sori and four clones that did not. We picked up the sorus contents of six random fruiting bodies from each and then spotted these individually on nutrient agar plates to test for bacterial growth. This assay confirmed that sori from some clones consistently contained bacteria that could initiate new populations (Fig. 1c, top panel) and that sori from other clones did not (Fig. 1c, bottom panel). Bacteria also grew when directional light induced the multicellular slugs to migrate away from the original locations to bacteria-free zones of a plate before fruiting, indicating that bacteria are carried in the slugs. Four sets of wild isolates tested in this way yielded 36% farmers: 13 of 35 tested overall; 4 of 9 from Minnesota; 5 of 14, 3 of 9 and 1 of 3 from three sample dates from Virginia. These data suggest that farmer clones are common, are found in the same locations as non-farmer clones and are therefore likely to have access to similar bacteria.

To confirm that farmers and non-farmers belong to the same species, we constructed a Bayesian phylogenetic tree created from combined mitochondrial, ribosomal and variable nuclear-DNA sequence data

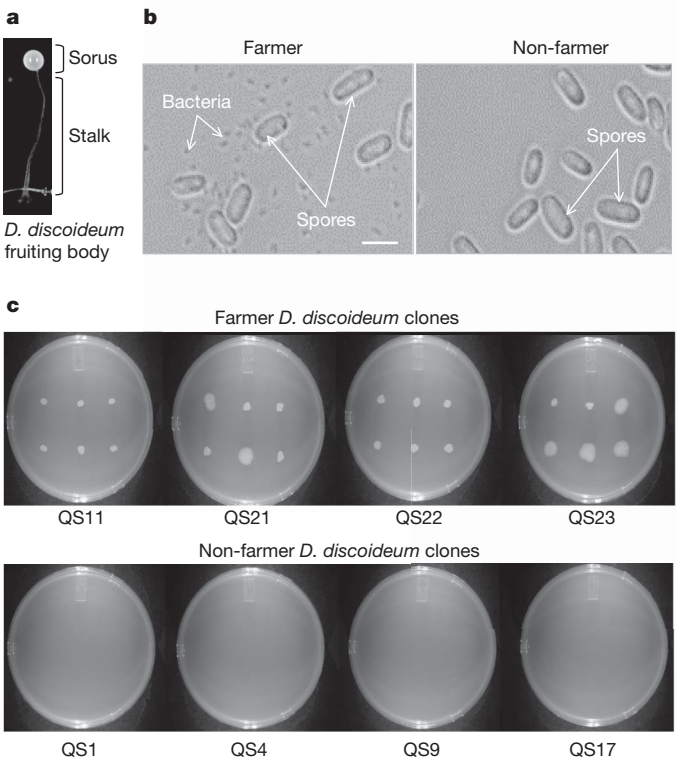


Figure 1 | Fruiting body structure and sorus contents from farmer and non-farmer *D.discoideum* clones. **a**, The fruiting body is composed of two main parts: the sorus, which contains the fertile spores, and the stalk, which holds the sorus aloft to facilitate spore dispersal. **b**, Bacteria and *D. discoideum* spores are present in the contents of a farmer sorus (left) but only spores are present in the contents of a non-farmer sorus (right). Scale bar, 5 μm. **c**, Sorus contents for six random, individual fruiting bodies from each of four farmers and four non-farmers were spotted on individual nutrient agar plates. All farmer sorus contents showed bacterial growth, whereas no bacterial growth was observed in sori from non-farmers. Plates were photographed after two days of growth.

[1]Department of Ecology and Evolutionary Biology, Rice University, 6100 Main Street, Houston, Texas 77005, USA.

from 14 clones (five farmers and nine non-farmers; Supplementary Fig. 1). Farmers were interspersed in the phylogeny with non-farmers, as expected if the trait is shuffled through the species by sex[22]. We also calculated pairwise genetic distances between farmer–farmer pairs, non-farmer–non-farmer pairs and farmer–non-farmer pairs and found no differences in the distributions (Supplementary Fig. 2).

We sequenced a portion of the bacterial 16S ribosomal gene to identify the species of carried bacteria. Farmers carry a variety of species of bacteria, with diversity both within and between farmer clones, and diversity is likely to be underestimated because not all bacteria are culturable by our methods (Supplementary Table 2). About half of these bacterial strains serve as good food sources for *D. discoideum*, generally for farmers and non-farmers alike (Supplementary Table 2). The function of the other half, if any, is unknown, but all farmer strains transport and use the food bacteria supplied in the lab (either *Klebsiella aerogenes* or *Escherichia coli*), and we focus the remainder of the paper on food carrying.

Carrying bacteria is a consistent clone-specific trait, and these clones show a number of differences from non-farmer clones that affect life history and fitness. To establish the consistency of the trait, we eliminated all living, carried bacteria from four farmers and four non-farmer controls by treating them with antibiotics. We then grew them on dead *K. aerogenes* as a food source, and confirmed using spotting tests as in Fig. 1 that they became bacteria free. When these bacteria-free clones were then grown on live *K. aerogenes*, all farmers regained an association and had these bacteria in their sori whereas non-farmers did not (Fig. 2).

To examine costs and benefits in the farmer–bacteria interaction, we compared farmers with non-farmers under several conditions. Soil is a very structured, heterogeneous environment in the wild, where bacteria could occur in patchy, monospecific colonies of variable cell numbers[23,24], and the patchiness is accentuated because some dictyostelids have distinct food preferences[25]. We therefore mimicked spore dispersal to both bacteria-poor and bacteria-rich sites. For the former, we transferred *D. discoideum* spores to plates with nutrients for bacteria, but without added bacteria (Fig. 3a). We found that bacteria transported by the farmer spores proliferate and are consumed by the farmers that also proliferate and then sporulate after the social stage, whereas non-farmers with no bacterial partners produce hardly any spores ($F_{1,6} = 58.97$ (derived from F-test; subscripts, degrees of freedom), $P < 0.0001$). Farmers are thus able to capitalize on available nutrients by carrying their own food bacteria in their sori. This difference disappears, and farmers do as poorly as non-farmers, if they are previously made bacteria free using antibiotics ($F_{1,6} = 0.49$, $P = 0.8082$; data not

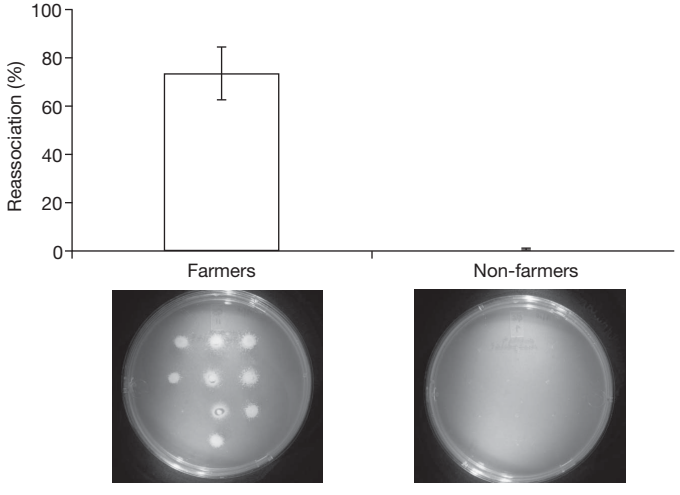


Figure 2 | Farmers readily reassociate with bacteria, suggesting a persistent interaction. After antibiotic treatment to eliminate carried bacteria, reassociation with bacteria was tested by growing four farmer and four non-farmer clones on *K. aerogenes*. We spotted ten random sori from each clone individually on nutrient agar plates, and counted positive/negative growth of bacteria in each sorus after two days. Farmers significantly differ from non-farmers in their ability to associate with bacteria ($F_{1,6} = 48.864$, $P < 0.001$; error bars, s.e.m.). Pictured below the graph are representative examples of sorus contents for farmer and non-farmer clones after one round of growth with live bacteria. Farmer clone sori contain bacteria; no bacterial growth was detected for non-farmers, even after ten days.

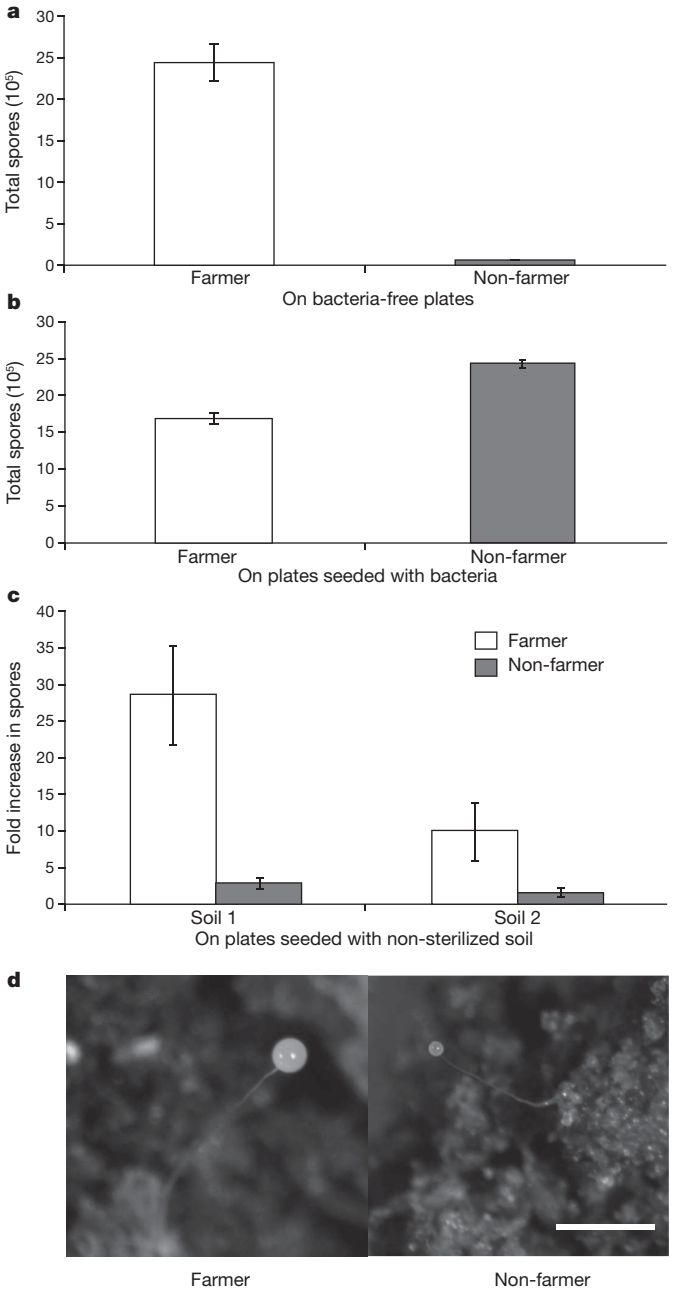


Figure 3 | The advantage of carrying food is context dependent. **a**, Farmers produce more spores than non-farmers when colonizing plates with bacterial food substrates but without bacteria ($F_{1,6} = 58.97$, $P < 0.0001$). We used three replicates of each of four farmer and four non-farmer clones. **b**, Farmers produce fewer spores than non-farmers when provided a fixed amount of live bacteria ($F_{1,12} = 64.36$, $P < 0.0001$). We used three replicates from each of five farmer and nine non-farmer clones (the results are similar when only the eight clones in **a** are used). **c**, Farmers produce more spores than non-farmers in unsterilized soil (soil 1: $F_{1,10} = 14.82$, $P = 0.0032$; soil 2: $F_{1,10} = 26.21$, $P = 0.0005$). We used individual sori collected from six farmers and six non-farmers to test spore production after one round of the social stage. All error bars, s.e.m. **d**, Representative examples of farmer and non-farmer fruiting bodies formed after the social stage in unsterilized soil. Scale bar, 1 mm.

shown). Next we tested whether farming was costly when spores are transferred instead to a site where edible bacteria are abundant, reducing or eliminating the advantage to farmers of bringing their own bacteria. Farmer clones then produce fewer spores than non-farmers from a given number of live bacteria ($F_{1,12} = 64.36$, $P < 0.0001$; Fig. 3b). Therefore, carrying seed stocks can be advantageous or disadvantageous, depending on bacterial availability at new sites.

Sites entirely lacking bacteria in nature are rare, but farmers could still gain by bringing preferred bacteria (whatever allowed them to flourish and sporulate in their previous site), just as humans seed preferred plants in an already green world. We therefore tested spore production for farmers and non-farmers in unsterilized soil collected from two separate locations (soil 1 and soil 2) at the Houston Arboretum and Nature Center, Texas. We determined the numbers of colony-forming units of culturable bacteria in soil 1 and soil 2 to be $(1.3–2.3) \times 10^8$ and $(0.6–0.64) \times 10^8$ per gram of soil, respectively. However, the bacteria already present in the soil do not make bacteria carrying superfluous. Under these conditions, farmers produced more spores than non-farmers for both soil locations (soil 1: $F_{1,10} = 14.82$, $P = 0.0032$; soil 2: $F_{1,10} = 26.21$, $P = 0.0005$; Fig. 3c, d), as well as many more fruiting bodies (soil 1: $F_{1,10} = 35.78$, $P = 0.0001$; soil 2: $F_{1,10} = 9.31$, $P = 0.0122$; data not shown). Farmer sori continued to carry their original bacteria (43 of 44 isolates checked by sequencing) whereas no bacteria were isolated from non-farmer sori. This suggests the bacteria available in the two test soils were not very suitable for both non-farmers and farmers, so the bacteria carried by the farmers allowed them to flourish in comparison with the non-farmers.

We proposed that the lower success of farmers when bacteria are provided stems from prudent harvesting. *D. discoideum* amoebae normally leave the solitary stage and enter the social stage when food is exhausted[16,17], but farmers may do so sooner to save some bacteria for transport. We therefore measured the number of uneaten bacteria present along a developmental time course (Fig. 4a). During the solitary amoeba stage (day 1), there was no difference in bacterial density among treatments (also, day-1 farmers and non-farmers did not differ significantly in number of amoebas: $F_{1,17} = 0.6733$, $P = 0.4233$; data not shown). During the social stage, however, bacterial usage differs significantly between farmers and non-farmers at all time points. Non-farmers eat all the bacteria whereas farmers leave many bacteria unconsumed, roughly half the number present as compared with bacteria grown alone. Figure 4b shows representative farmer and non-farmer clones photographed at day 5, revealing that only the farmers entered the social stage and formed fruiting bodies in the presence of uneaten bacteria. Thus, it seems that farmer clones forgo considerable food to save some for co-dispersal.

Farmer clones also migrate significant less far than non-farmers during the mobile slug stage that immediately precedes fruiting ($F_{1,14} = 87.59$, $P < 0.001$; Fig. 4c). This might be a cost caused by bacterial interference, or it might be an evolved response of not needing to move as far when farmers carry their own bacteria. Either way, it adds to the list of significant differences between farmers and non-farmers.

An alternative explanation for the apparent costs—leaving some bacteria unharvested and reduced slug migration—is that the bacteria are pathogenic and harm *D. discoideum*. However, the pathogen hypothesis does not account for why the farmer strains would be infectible by many bacteria and why infection is highly consistent, even after curing. Neither does it explain why all farmers carry food. Moreover, it does not explain why infection causes no cost to growth in the solitary stage ($F_{1,11} = 0.72$, $P = 0.4132$; Supplementary Fig. 3). Instead, the costs appear precisely where the farming hypothesis predicts (saving some food for transport) or where it provides a plausible explanation (less need for slug migration). However, a mixed explanation seems possible. Carrying food bacteria could have the side effect of sometimes taking up useless or harmful bacteria. The existence of the farmer polymorphism will allow additional within-species comparisons that should help in exploring the mechanisms, which are as yet unknown,

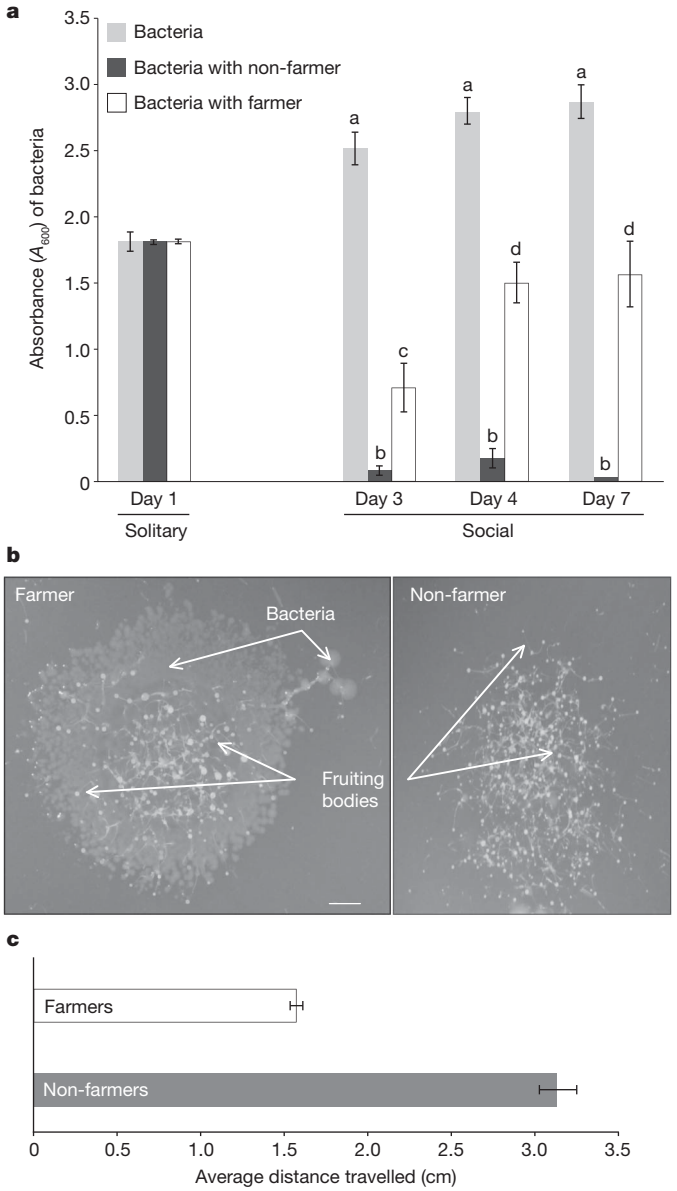


Figure 4 | Life-history traits differ between farmers and non-farmers. **a**, Bacterial usage. For 11 non-farmer clones and nine farmer clones, we individually spotted a fixed number of spores mixed with the food bacterium *K. aerogenes* on nutrient agar plates. We also spotted *K. aerogenes* alone as a control. Bacterial density for each spot was determined at various time points during development. During the social stage (days 3, 5, 7), farmers leave more bacteria unconsumed than non-farmers at all time points (type: $F_{2,19} = 106.18$, $P < 0.0001$; day: $F_{2,38} = 10.23$, $P = 0.0003$; type × day: $F_{4,38} = 10.60$, $P < 0.0001$; significant differences found within each day are indicated by different letters, which reflect results of a *post hoc* Tukey HSD test). At the solitary-amoeba stage (day 1), no difference in bacterial density was found among the three treatments ($F_{2,19} = 1.2943$, $P = 0.2972$). **b**, Bacterial usage of representative clones. Farmer clones develop fruiting bodies before all bacteria are exhausted whereas non-farmers consume all accessible food sources before fruiting. Examples of fruiting bodies are marked with arrows and appear as white dots. Clones were photographed on day 5. Scale bar, 3 mm. **c**, Migration. The average distance farmer slugs migrate towards light is less than for non-farmer slugs ($F_{1,14} = 87.59$, $P < 0.001$). Eight clones of each type were used, with two replicates. All error bars, s.e.m.

but could be very simple alterations in timing of aggregation, sensitivity to bacteria or ability to produce a specific enzyme or toxin. The abundant scientific resources available for this model organism have recently proven very useful in understanding the genetics of their social interactions[26–29]. They should prove similarly useful here, providing a

unique model system for probing the genetics of eukaryotic–bacterial symbioses.

The connection between farming and sociality may not be coincidental, because social species have suitably structured populations. In this social microbe, the advantage of prudent harvesting and seeding is large because it can benefit many generations of cell descendants before fruiting. Moreover, the high relatedness of natural fruiting bodies[27] minimizes any potential exploitation by non-farmers, which could either consume the bacteria that the farmers would save to carry, or freeload when co-dispersed with farmer spores and their bacteria. This same advantage—long-lived groups of kin—provides similarly fertile ground for agriculture in the ants and termites that are the most sophisticated non-human farmers.

METHODS SUMMARY

Isolation of wild *D. discoideum* strains. Isolation techniques followed published methods[30] with the exception that we collected soil samples of 20 g or more, instead of 0.2 g, at each location.

Culture conditions. We grew all wild isolates from spores on SM/5 agar plates (2 g glucose, 2 g BactoPeptone (Oxoid), 2 g yeast extract (Oxoid), 0.2 g $MgCl_2$, 1.9 g $KHPO_4$, 1 g K_2HPO_4 and 15 g agar per litre) in association with bacteria *K. aerogenes* or *E. coli* at room temperature (21 °C).

PCR amplification and sequence identification of novel bacterial isolates. We followed the procedures outlined in "Identifying unknown bacteria using biochemical and molecular methods" (http://www.nslc.wustl.edu/elgin/genomics/Bio3055/IdUnknBacteria06.pdf) with one exception. Bacteria to be cloned and identified were grown on and collected from SM/5 agar plates. PCR fragments generated (using the above procedure) were sequenced at Lone Star Labs (Houston, Texas). We used the NCBI website http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi as the search tool for sequences to identify bacteria to species.

Data analyses. We analysed our data using standard analysis-of-variance methodology with fixed (farmer and non-farmer) effects and a random effect (clone) for all experimental assays. The data analysis was generated using SAS software (version 9-1 of the SAS System for Microsoft Windows).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 28 April; accepted 12 November 2010.

1. Smith, B. D. (ed.) *The Emergence of Agriculture* (Scientific American Library, Freeman, 1995).
2. Aanen, D. *et al.* The evolution of fungus-growing termites and their mutualistic fungal symbionts. *Proc. Natl Acad. Sci. USA* **99,** 14887–14892 (2002).
3. Farrell, B. *et al.* The evolution of agriculture in beetles (Curculionidae: Scolytinae and Platypodinae). *Evolution* **55,** 2011–2027 (2001).
4. Mueller, U. G., Schultz, T., Currie, C., Adams, R. & Malloch, D. The origin of the ant-fungus mutualism. *Q. Rev. Biol.* **76,** 169–197 (2001).
5. Mueller, U. G., Gerardo, N. M., Aanen, D. K., Six, D. L. & Schultz, T. R. The evolution of agriculture in insects. *Annu. Rev. Ecol. Evol. Syst.* **36,** 563–595 (2005).
6. Silliman, B. R. & Newell, S. Y. Fungal farming in a snail. *Proc. Natl Acad. Sci. USA* **100,** 15643–15648 (2003).
7. Hata, H. & Kato, M. A novel obligate cultivation mutualism between damselfish and Polysiphonia algae. *Biol. Lett.* **2,** 593–596 (2006).
8. Crespi, B. J. The evolution of social behavior in microorganisms. *Trends Ecol. Evol.* **16,** 178–183 (2001).
9. Keller, L. & Surette, M. G. Communication in bacteria: an ecological and evolutionary perspective. *Nature Rev. Microbiol.* **4,** 249–258 (2006).
10. Mehdiabadi, N. J. *et al.* Kin preference in a social amoeba. *Nature* **442,** 881–882 (2006).
11. Ostrowski, E. A., Katoh, M., Shaulsky, G., Queller, D. C. & Strassmann, J. E. Kin discrimination increases with genetic distance in a social amoeba. *PLoS Biol.* **6,** e287 (2008).
12. Douglas, A. E. Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria *Buchnera*. *Annu. Rev. Entomol.* **43,** 17–37 (1998).
13. Moran, N. A., Dunbar, H. E. & Wilcox, J. L. Regulation of transcription in a reduced bacterial genome: nutrient-provisioning genes of the obligate symbiont Buchnera aphidicola. *J. Bacteriol.* **187,** 4229–4237 (2005).
14. Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. Host-bacterial mutualism in the human intestine. *Science* **307,** 1915–1920 (2005).
15. Sears, C. L. A dynamic partnership: celebrating our gut flora. *Anaerobe* **11,** 247–251 (2005).
16. Raper, K. B. *The Dictyostelids* 87–177 (Princeton Univ. Press, 1984).
17. Kessin, R. H. *Dictyostelium: Evolution, Cell Biology, and the Development of Multicellularity* (Cambridge Univ. Press, 2001).
18. Raper, K. B. Growth and development of Dictyostelium discoideum with different bacterial associates. *J. Agric. Res.* **55,** 289–316 (1937).
19. Matz, C. & Kjelleberg, S. Off the hook - how bacteria survive protozoan grazing. *Trends Microbiol.* **13,** 302–307 (2005).
20. Nadson, G. A. Des cultures du Dictyostelium mucoroides Bref. et des cultures pures des amibes en general. *Scripta Bot. Horti Univ. Imp. Petropolitanae* **15,** 188–190 (1899).
21. Skupienski, F. X. *Recherches sur le Cycle Evolutif de Certains Myxomycetes*. PhD thesis, l'Universite de Paris (1920).
22. Flowers, J. M. *et al.* Variation, sex, and social cooperation: molecular population genetics of the social amoeba *Dictyostelium discoideum*. *PLoS Genet.* **6,** e1001013 (2010).
23. Clark, F. in *Soil Biology* (eds Burges, A. & Raw, F.) Ch. 2, 15–49 (Academic, 1967).
24. Heijnen, C. E., Burgers, S. & Vanveen, J. A. Metabolic activity and population dynamics of rhizobia introduced into unamended and bentonite-amended loamy sand. *Appl. Environ. Microbiol.* **59,** 743–747 (1993).
25. Horn, E. G. Food competition among the cellular slime molds (Acrasiae). *Ecology* **52,** 475–484 (1971).
26. Foster, K. R., Shaulsky, G., Strassmann, J. E., Queller, D. C. & Thompson, C. R. L. Pleiotropy as a mechanism to stabilize cooperation. *Nature* **431,** 693–696 (2004).
27. Gilbert, O. M., Foster, K. R., Mehdiabadi, N. J., Strassmann, J. E. & Queller, D. C. High relatedness maintains multicellular cooperation in a social amoeba by controlling cheater mutants. *Proc. Natl Acad. Sci. USA* **104,** 8913–8917 (2007).
28. Santorelli, L. A. *et al.* Facultative cheater mutants reveal the genetic complexity of cooperation in social amoebae. *Nature* **451,** 1107–1110 (2008).
29. Benabentos, R. *et al.* Polymorphic members of the *lag* gene family mediate kin discrimination in *Dictyostelium*. *Curr. Biol.* **19,** 567–572 (2009).
30. Fortunato, A., Strassmann, J. E., Santorelli, L. & Queller, D. C. Co-occurrence in nature of different clones of the social amoeba, Dictyostelium discoideum. *Mol. Ecol.* **12,** 1031–1038 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Rudgers, G. Saxer Quance, L. Campbell, E. Ostrowski, O. Gilbert, A. Savage, J. Ahern, K. Crawford, S. Chamberlain, S. Read, D. Nguyen, K. Foster, H. Kaplan, D. Hatton and K. Boomsma for discussions and advice. This material is based on work supported by the US National Science Foundation.

Author Contributions D.A.B. identified the symbiosis, performed the experiments and analysed the data. T.E.D. constructed and analysed the phylogeny. D.A.B., T.E.D., D.C.Q. and J.E.S. designed the experiments, discussed the results and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.A.B. (dbrock@rice.edu).

METHODS

Isolation of wild *D. discoideum* strains. Isolation techniques followed published methods[30] with the exception that we collected soil samples of 20 g or more, instead of 0.2 g, at each location.

Culture conditions. We grew all wild isolates from spores on SM/5 agar plates (2 g glucose, 2 g BactoPeptone (Oxoid), 2 g yeast extract (Oxoid), 0.2 g $MgCl_2$, 1.9 g $KHPO_4$, 1 g K_2HPO_4 and 15 g agar per litre) in association with bacteria *K. aerogenes* or *E. coli* at room temperature.

PCR amplification and sequence identification of novel bacterial isolates. We followed the procedures outlined in "Identifying unknown bacteria using biochemical and molecular methods" (http://www.nslc.wustl.edu/elgin/genomics/Bio3055/IdUnknBacteria06.pdf) with one exception. Bacteria to be cloned and identified were grown on and collected from SM/5 agar plates. PCR fragments generated (using the above procedure) were sequenced at Lone Star Labs (Houston, Texas). We used the NCBI website http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi as the search tool for sequences to identify bacteria to species.

Data analyses. We analysed our data using standard analysis-of-variance methodology with fixed (farmer and non-farmer) effects and a random effect (clone) for all experimental assays. The data analysis was generated using SAS software (version 9-1 of the SAS System for Microsoft Windows).

Experimental assays. Farmer test. We picked up the sorus contents of 6–12 random fruiting bodies developed on SM/5 plates in association with *E. coli* or *K. aerogenes* from each wild clone to be tested using a filtered pipette tip. The sorus contents were spotted individually on SM/5 agar plates and assessed for bacterial growth after two days at room temperature.

Reassociation assay. We used a population of four farmers and four non-farmers treated to eliminate all living, carried bacteria. To accomplish this, clones were grown on SM/5 agar plates containing antibiotics (0.1 g ampicillin, 0.3 g streptomycin sulphate per litre) using dead *E. coli* as a limited food source to allow the spores to hatch and develop. Dead *E. coli* was prepared by autoclaving a suspension of *E. coli* and KK2 (2.25 g KH_2PO_4 and 0.67 g K_2HPO_4 per litre H_2O) for 20 min at 121 °C. After autoclaving, the absorbance of dead bacteria was set to A_{600} 6.0. Treated clones were allowed to form fruiting bodies, spores were collected and the process was repeated. We performed the farmer test after the second round and no bacterial growth was seen in the sorus contents of test clones even after ten days. We harvested spores from the bacteria-free, cured clones in KK2 and the spore density for each clone was determined by serial dilution using a haemacytometer under a light microscope. Spores from these treated clones were then reintroduced to live *E. coli*, plated on SM/5 agar plates, and allowed to form fruiting bodies. Ten random sori were collected individually from each clone and tested for bacterial growth following the farmer test above.

Migration assay. We tested the slug migration ability of eight farmers and eight non-farmers per replicate, for two replicates. Duplicate plates were set up for each clone per replicate. For each plate, we prepared a slurry of 5×10^6 spores of one *D. discoideum* clone, *K. aerogenes* and KK2. We then applied the slurry to the edge of a $150 \times 15\,mm^2$ Petri plate containing 80 ml nutrient-free agar (0.198 g KH_2PO_4, 0.0356 g Na_2HPO_4 and 15 g agar per litre), and allowed the slurry to dry. Each plate was wrapped in aluminium foil and a small pinhole was made directly opposite the spores and bacteria to provide directional light. The wrapped plates were placed in a lit incubator at 21 °C for 84 h to allow slugs to form and migrate. For each clone, we counted the distance travelled by slugs in each of five 1.5-cm bands across the nutrient-free agar plate, and calculated the average distance travelled for all slugs.

Spore production assay. We tested two conditions: first, nutrients for bacterial growth but no outside bacteria were provided as food for the test clones; second, nutrients for bacterial growth as well as outside bacteria were provided as food for the test clones. For the first condition, we used four farmer clones and four non-farmer clones per replicate, with three replicates. We spread 200 µl of 10^5 spores plus dead *E. coli* in KK2 (absorbance, A_{600} 6.0) on $100 \times 15\,mm^2$ SM/5 agar plates. Bacterial growth is possible for farmers in this condition. After development was complete, spores were collected by washing the plates with KK2 plus 0.1% NP-40 (Calbiochem). The total number of spores produced by each clone was determined by counting using a haemacytometer and a light microscope. As a control to assess confounding growth differences between farmers and non-farmers, all eight clones were grown as above but the plates were supplemented with antibiotics (see "Reassociation assay", above) to eliminate any potential food bacteria carried by the clones. No farming is possible in this set-up.

For the second condition, we used five farmer clones and nine non-farmer clones per replicate, with three replicates. For the assay, we prepared nutrient-free agar plates (see "Migration assay", above) laid with a grid of equidistant 13-mm AABP 04700 (Millipore) black filter squares. Filters were spotted individually with 5×10^5 spores in a slurry of live *K. aerogenes* and KK2 (A_{600} 6.0). Clones were spotted in an order determined by a random number generator. Duplicate samples were made for each clone for each experiment. Clones were allowed to hatch, grow and develop under direct light to limit potential movement of slugs before final culmination in fruiting bodies. Development was complete for all clones after four days. Each filter was collected and placed in a 1.5-ml conical Eppendorf tube containing 1.0 ml KK2 plus 0.1% NP-40 alternative. Tubes were vortexed briefly to disperse the spores evenly and counted as above without dilution to determine density.

Soil assay. We collected soil from two separate locations in the Houston Arboretum and Nature Center (29° 46′ N, 95° 27′ W). Thin, non-nutrient agar plates were prepared to provide a humid environment for spores to hatch and to hold the soil in place. Soil was then layered to a ~2–3-mm depth in a half-moon pattern with an empty space at the centre line atop the starving agar on each plate (Supplementary Fig. 4a). We used 12 clones consisting of six farmers and six non-farmers. Farmer–non-farmer clones were randomly paired on each plate (one on each half-moon of soil) as a check for plate environment bias. Each pair was tested on two separate plates for each soil type. Data analysis was reported unpaired as no plate environment bias was detected.

For the assay, fruiting body sori were collected from stock plates of all 12 clones previously prepared on the same day by plating 2×10^5 spores in association with *K. aerogenes* on SM/5 agar plates. The contents of an individual sorus for either a farmer or a non-farmer clone were picked up using a filtered pipette tip and placed on the unsterilized soil in the previously chosen locations (marked by coloured circles in Supplementary Fig. 4b). Three farmer sori and three non-farmer sori were placed on each experimental plate in this manner. After three days at room temperature under direct overhead light, all plates were viewed under a dissecting microscope to locate and collect all fruiting bodies formed. Fruiting bodies found were generally located in the same area as spotted on the soil. Whole fruiting bodies from all three spots per plate for either the farmer or the non-farmer were collected and placed together in an Eppendorf tube containing 1 ml KK2 plus 0.1% NP-40. Tubes were vortexed briefly to disperse the spores, and the spores were counted without dilution using a haemacytometer. The change in spore number for each clone was then calculated. To determine the initial number of spores spotted for each clone without diluting sorus contents, a proxy was used. The sorus contents from ten random fruiting bodies from each clone were collected and counted as above. The average count for these ten fruiting bodies was used to determine the clone's average spore number per sorus spotted on the experimental plates. Additionally, we determined presence or absence of bacteria in the sori as well as bacteria identity for positive growth using a subset of ten clones (five farmers and five non-farmers) in both types of soil. We performed serial dilutions in KK2 of 80 sori contents (44 farmer and 34 non-farmer), and the clonal isolates recovered were used for sequencing following the methods described in "PCR amplification and sequence identification of novel bacterial isolates", above.

Bacteria usage assay. We collected spores from a population of 11 non-farmer and nine farmer clones. For each clone, we individually spotted four 30-µl spots of 3×10^4 spores mixed with live *K. aerogenes* (A_{600} 3.0) on a SM/5 plate as well as spotting *K. aerogenes* alone as a control. Bacterial density was determined by using a sterilized inoculating loop to collect all growth from one spot for each clone in a 1.5-ml Eppendorf tube containing 1 ml KK2, vortexing to obtain a uniform suspension, removing hatched amoeba or spores by briefly centrifuging at 2,000*g* to pellet, and then determining the absorbance (A_{600}) of the remaining bacteria. Data points were collected on days 1, 3, 5 and 7. To determine whether confounding differences in spore germination occurred among clones, the number of hatched amoeba was determined for each clone on day 1 by counting using a haemacytometer.

Proliferation assay. To determine vegetative doubling rates during exponential growth, we grew each clone separately by plating 1×10^4 log-phase cells per plate in association with *K. aerogenes* as a food source on replicated SM/5 agar plates. After 12 h of growth, we collected all cells from a plate, diluted the cells in a measured amount of KK2 and counted the number of cells present using a haemacytometer. We repeated this process for plates grown for 18, 24, 30, 36 and 42 h. We conducted the experiment in two temporally independent blocks. To analyse the data, we log-transformed the counts, determined the slope for each clone and performed a full-factor analysis of covariance.

Phylogeny construction. DNA sequencing. We extracted DNA from spores using a Chelex/proteinase K extraction protocol. We amplified a non-coding region of the mitochondrial genome (mtDNA), regions of the nuclear ribosomal DNA (rDNA) and six variable fragments of nuclear DNA by PCR with the primers listed in Supplementary Table 3, using the following protocol (step 1: 2 min at 94 °C; step 2: 30 s at 94 °C; step 3: 30 s starting at 65 °C and decreasing by 1 °C per cycle; step 4: 1 min at 72 °C; step 5: 15 cycles to step 2; step 6: 30 s at 94 °C; step 7: 30 s at 50 °C; step 8: 1 min at 72 °C; step 9: 25 times to step 6; step 10: 6 min at 72 °C). We cleaned the PCR product with USB ExoSAP-IT and then sequenced using Perkin Elmer Applied Biosystems Big Dye 3.1 chemistry and a 3100 genetic

analyser. We analysed approximately 4,300 base pairs of the nucleotide sequence of nuclear 17S, 5.8S, 26S and 5S rDNA regions, approximately 800 base pairs of the nucleotide sequence of mtDNA (LSU intron) and approximately 3,700 base pairs of variable nuclear DNA from chromosomes 1, 2, 3 and 4. We aligned the sequences using the programs LASERGENE SEQMAN (version 7.0.0) and BIOEDIT (version 7.0.5.2).

Data analysis. We used comparative DNA sequence data from 14 individual clones (QS numbers 1, 4, 6, 8-9, 11-12, 14-15, 17-18 and 21-23) to estimate gene trees/phylogenies and to estimate pairwise genetic distances between clones. These 14 clones represent five farmers and nine non-farmers. We used Bayesian methods for phylogenetic reconstruction. Using MRBAYES[31] (version 3.1), we estimated a phylogeny for each data set based on the GTR+I+Γ model of molecular evolution. In addition, two high-frequency, polymorphic indels (one in the mtDNA and one in the variable nuclear DNA) were scored as standard presence/absence characters and were included in the analysis with weighting equal to the nucleotide polymorphisms. For each analysis, four Metropolis-coupled Markov chains were run for 250,000 burn-in generations followed by 1.75×10^6 generations of data collection. We used the software program MEGA[32] (version 4) to estimate pairwise genetic distances between clones using the *p*-distance algorithm. We analysed sequence data from all 14 individual clones. Gaps and missing data were eliminated in pairwise sequence comparisons.

31. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17,** 754–755 (2001).
32. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24,** 1596–1599 (2007).

LETTER

doi:10.1038/nature09633

Preplay of future place cell sequences by hippocampal cellular assemblies

George Dragoi[1] & Susumu Tonegawa[1]

During spatial exploration, hippocampal neurons show a sequential firing pattern in which individual neurons fire specifically at particular locations along the animal's trajectory (place cells[1,2]). According to the dominant model of hippocampal cell assembly activity, place cell firing order is established for the first time during exploration, to encode the spatial experience, and is subsequently replayed during rest[3–6] or slow-wave sleep[7–10] for consolidation of the encoded experience[11,12]. Here we report that temporal sequences of firing of place cells expressed during a novel spatial experience occurred on a significant number of occasions during the resting or sleeping period preceding the experience. This phenomenon, which is called preplay, occurred in disjunction with sequences of replay of a familiar experience. These results suggest that internal neuronal dynamics during resting or sleep organize hippocampal cellular assemblies[13–15] into temporal sequences that contribute to the encoding of a related novel experience occurring in the future.

We recorded neuronal firing sequences from the CA1 area of the mouse hippocampus (Supplementary Fig. 1) during periods of awake rest (Fam-Rest) alternating with periods of running (Fam-Run) on a familiar track (Fam session; Supplementary Fig. 2a) that preceded the exploration of a novel linear arm in contiguity with the familiar track (Contig-Run on L-shaped track; Fig. 1, Supplementary Fig. 2a and Methods). All the place cells active on the novel arm during Contig-Run, whether previously silent[16] (19% in both directions and 31% in at least one direction; Methods and Supplementary Tables 1–3) or active during Fam-Run (subpanels a in Fig. 1), fired during Fam-Rest at the ends of the familiar track (range, 0.17–11.7 Hz; Supplementary Fig. 3) as part of a number of 'spiking events'. The spiking events were defined as epochs composed of multiple individual spikes from at least four different place cells active on the novel arm or familiar track, separated by less than 50 ms and flanked by at least 50 ms of silence[3,4]. More significantly, the temporal sequence in which the cells active on the novel arm fired during Fam-Rest (subpanels b in Fig. 1) was significantly correlated with the spatial sequence in which they fired later as place cells on the novel arm during Contig-Run (subpanels c in Fig. 1), despite being uncorrelated with their spatial sequence as place cells on the familiar track during Fam-Run. This is illustrated as place cell sequences during Contig-Run (subpanels c in Fig. 1) and Fam-Run (subpanels a in Fig. 1) compared with the firing sequences of these cells within individual spiking events observed during Fam-Rest (subpanels b in Fig. 1). We refer to this process as 'preplay' of place cell sequences because the temporal sequence of firing during Fam-Rest had occurred before the actual exploration of the novel arm in the subsequent Contig-Run and was not a replay of the place cell sequences from the previous Fam-Run.

To quantify the significance of preplay and to compare it with replay, we created place cell sequence templates according to the spatial order of the peak firing of place cells[3,4,10] on the novel arm during Contig-Run (novel arm templates; subpanels c in Fig. 1 and Methods) and on the familiar track during Fam-Run (familiar track templates) for each run direction. The spikes of all the place cells used to construct the two types of template that were emitted during Fam-Rest were sorted by time, and spiking events were determined as explained above (subpanels b in Fig. 1). For each spiking event, we calculated a rank-order correlation between the novel arm templates and the temporal sequence of firing of the corresponding cells in the spiking events during Fam-Rest. The event correlation was considered significant if it exceeded the 97.5th percentile of a distribution of correlations resulting from randomly shuffling the order of place cells in the novel arm templates 200 times ($P < 0.025$). Forward[4] and reverse[3,4] preplay refers to the cases in which the sequence of place cells during Contig-Run and the firing order of the corresponding cells in Fam-Rest were in the same and opposite directions, respectively. In 91% of the preplay cases, the spiking events were correlated with the novel arm template in one direction only. The distribution of event correlation values obtained using the original novel arm templates was significantly shifted towards higher positive or negative values in comparison with the distribution of correlation values obtained using shuffled templates (Fig. 2a and Supplementary Fig. 4). Figure 2a also shows the distribution of significant preplay events (in red). Of all the spiking events detected as above and in which at least four novel arm place cells were active, 14.2% were significant preplay events for the place cell sequence on the novel arm ($P < 10^{-32}$, binomial probability test[4]) in the forward or reverse order (Fig. 2b).

The occurrence of significant preplay events was correlated with the occurrence of high-frequency ripple oscillations in CA1 (Fig. 2c). The majority of the significant preplay events (81.1%; Fig. 2d, total, blue) took place at the junction between the familiar and novel arms, and the remaining 18.9% took place at the free end of the familiar track (Fig. 2d, total, purple). The proportion of significant preplay events among the total events at each of the two track ends was higher at the junctional end (15.2%, $P < 10^{-26}$) than at the free end (8.5%, $P < 10^{-4}$) of the familiar track ($P < 0.035$, Z-test; Fig. 2d, normalized).

We found a relatively high correlation between the place field maps (Fig. 1A, B and Supplementary Fig. 5) of the familiar track before and after the novel experience (median $r = 0.66$; Fig. 2e, familiar track, blue); it was significantly higher than the correlations obtained when the cell identities were shuffled (median $r = 0.23$, $P < 10^{-4}$; Fig. 2e, familiar track, black). A similar correlation analysis showed a relatively high stability of the newly formed place fields on the novel arm from the beginning to the end of Contig-Run (median $r = 0.62$ (newly formed) versus median $r = 0.21$ (shuffled), $P < 10^{-3}$; Fig. 2e, novel arm, blue versus grey). These results suggest that preplay of the novel arm does not occur over an entirely new (that is, remapped) representation of the whole L-shaped track but rather benefits from the relative stability of the familiar track representation across sessions and perhaps facilitates the rapid, stable encoding of the novel arm experience.

Using the familiar track templates and spiking events during Fam-Rest, constructed as above, we determined that 16.2% ($P < 10^{-91}$; data not shown) were significant replay events[3–6,17] among the spiking events in which a minimum of four familiar track place cells were active. All significant preplay events occurring during Fam-Rest ($n = 75$) were

[1]The Picower Institute for Learning and Memory, RIKEN-MIT Center for Neural Circuit Genetics, Department of Biology and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

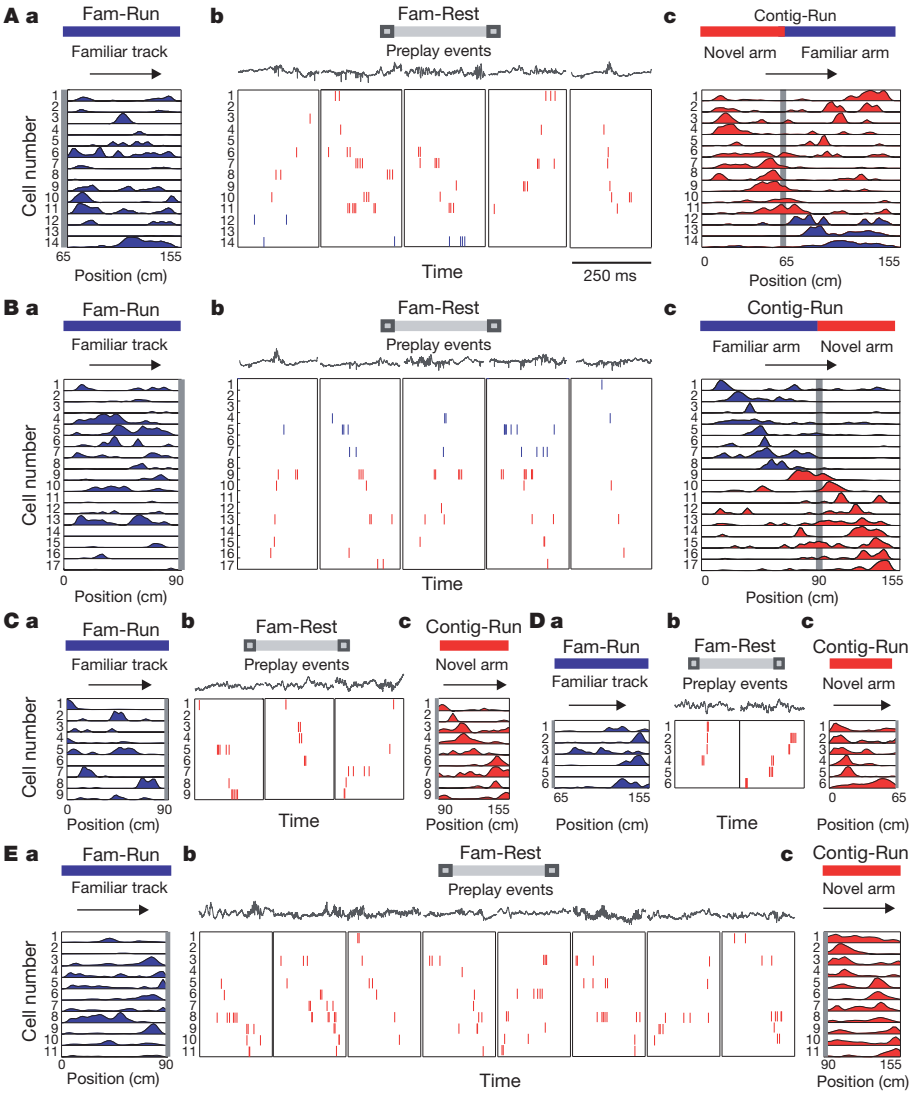


Figure 1 | Preplay of novel place cell sequences. Fam-Run and Fam-Rest respectively denote run and rest sessions on the familiar linear track before barrier removal; Contig-Run denotes run sessions on the L-shaped track after barrier removal. The L-shape track was linearized for display/analysis. **A**, **B**, mouse 1; **C**, **D**, mouse 2; **E**, mouse 3. **A**–**E**, **a**, Spatial activity on the familiar track during Fam-Run of the cells that had place fields in Contig-Run and preplayed during Fam-Rest (one cell per row); activity on the novel arm and familiar track are on the same scale. Horizontal arrows indicate run directions. Vertical grey bars indicate barrier locations during Fam-Run and Fam-Rest. **A**–**E**, **b**, Examples of representative spiking events in the forward or reverse direction during Fam-Rest in 250-ms time windows (350 ms for the second and fourth panels from left in **E**, **b**). Tick marks indicate individual spikes: red, preplay events for place cell sequences in the novel arm; blue (in **A**, **b** and **B**, **b**), additional spikes from the familiar track place cells participating in the spiking event (not shown in **C**–**E**, **b**). Numbers on the left denote cell numbers and correspond to the place cell numbers in **A**–**E**, **a**. Square boxes indicate the ends of the familiar track where preplay events occurred. Local field potentials recorded simultaneously with the spikes are shown above spiking events. **A**–**E**, **c**, Place cell sequences in the novel arm (**C**–**E**, **c**; red) or in both the novel arm (red) and the familiar arm (blue) (**A**, **c** and **B**, **c**) in Contig-Run.

tested for possible replay of the familiar track spatial sequence: these spiking events were more correlated with the novel arm template (Fig. 2f, red) than the familiar track template (Fig. 2f, blue). Seventy-two percent ($n = 54$) of the significant events previously considered to be preplay had no significant correlation with the familiar track template. An additional 16% ($n = 12$) of those events were better correlated with the novel arm templates (mean absolute $r = 0.92$) than with the familiar track template (mean absolute $r = 0.67$, $P < 10^{-3}$). Together, these findings reject the hypothesis that the preplay events simply represent a replay of the familiar track activity (see additional controls in Supplementary Information). Moreover, we found that the proportion of events exclusively composed of silent cells that perfectly matched the novel arm spatial templates was 0.67 (16 of 24 triplets), which is significantly greater ($P < 0.025$) than the proportion of by-chance perfect matches (0.33).

To illustrate the distribution and relative proportions of preplay and replay events among all significant spiking events during Fam-Rest, we calculated a 'template specificity index' (Fig. 2g and Methods) for each event. Pure preplay events (Fig. 2g, red) and pure replay events (Fig. 2g, blue) were segregated, and only a minority of events were significant for both preplay and replay (Fig. 2g, yellow). Consistent with this segregation of preplay and replay events, the novel arm and the 'corresponding familiar track' templates were not significantly correlated (Fig. 2h and Methods). The ratio between the number of pure replay events ($n = 171$) and the number of pure preplay events ($n = 54$) during Fam-Rest was about 3.1 (Fig. 2g, inset; see Supplementary Information for proportions of events). Preplay and replay events were distributed in time across Fam-Rest (Supplementary Fig. 6a–c) and their occurrences were generally uncorrelated (Supplementary Fig. 6d). The majority (79.9%) of the spiking events during Fam-Rest did not significantly correlate with either of the two templates (data not shown).

We used a Bayesian reconstruction algorithm[2,5,6,18,19] (Methods) to decode the animals' position from the spiking activity during Fam-Run (Fig. 3a) or Fam-Rest (Fig. 3b, c). For all original and shuffled[6]

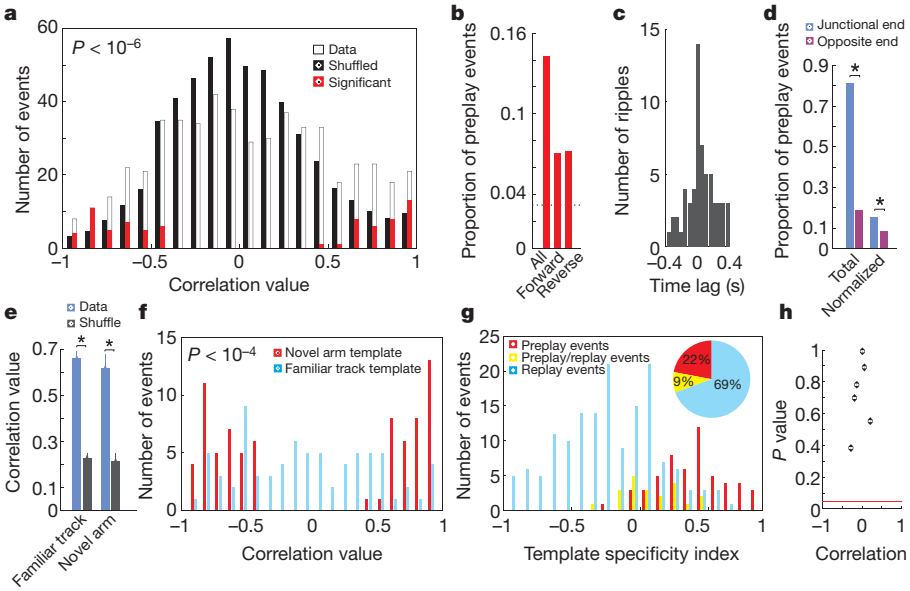


Figure 2 | Quantification of the preplay phenomenon and comparison with replay. **a**, Distribution of correlations between spiking events in Fam-Rest and spatial templates of the novel arm. Open bars indicate spiking events versus the original (unshuffled) templates; filled bars indicate spiking events versus 200 shuffled templates scaled down 200 times; red bars show the distribution of preplay (that is, significant) events. Similar distributions (not shown) of corresponding spiking events were obtained when spatial templates were constructed using all place cells active on the L-shaped track (Figs 1A, b, c and 1B, b, c; red and blue). **b**, Proportion of all, forward and reverse preplay events among the spiking events in Fam-Rest. The dotted line indicates the chance level (3.2%). **c**, Cross-correlation between preplay events and ripple epochs. **d**, Location of preplay events on the familiar track: total, proportions of preplay events at ends of the track; normalized, proportion of preplay events normalized by the number of spiking events at each end of track. Preplay events represented a trajectory running from the free end of the novel arm to the junctional end (40%) or begun near the familiar track (60%); the latter suggests that in some cases preplay events could be triggered by the activity of the familiar track place cells during Fam-Rest. **e**, Stability of place cell spatial tuning across the novel experience: familiar track, stability of the place fields active on the familiar track before (Fam-Run) versus after (Contig-Run) barrier removal; novel arm, stability of the place fields active on the novel arm at the beginning (first four laps of run) versus the end (last four laps) of the Contig-Run session. Data (blue), within-cell correlation of place cell spatial tuning for the corresponding track/arm; shuffle (black), cell identity shuffle (Supplementary Information). Error bars, s.e.m.; asterisks in **d** and **e** indicate significant differences. **f**, Distribution of preplay event correlations (red) versus distribution of these event correlations with the familiar track template (blue). Spiking events were detected using all place cells from the familiar track and novel arm templates (>1 Hz). Red bars are the same as in **a**. Correlation is strong with the novel arm template (preplay) and weak with the familiar arm template (replay). The *P* value corresponds to there being a significant difference between the two distributions. **g**, Disjunctive distribution of pure preplay (red), pure replay (blue) and preplay/replay (yellow) events during Fam-Rest over their template specificity index (Supplementary Information). Inset, proportions of pure preplay events (red), pure replay events (blue) and preplay/replay events (yellow) among all of the spiking events that were significantly correlated with at least familiar track templates or novel arm templates. **h**, Lack of correlation between the novel arm template and the corresponding familiar track template. Each of the six dots represents either a forward or a reverse run direction of one of the three mice analysed. Red horizontal line denotes a *P* value of 0.05. The correlation values were not significant in any of the cases (Supplementary Information).

probability distributions, a line was fitted to the data using a line-finding algorithm[6] to represent the decoded virtual trajectory (Methods and Supplementary Information). In 16.36% of cases representing trajectories, the reconstructed trajectory during spiking events in Fam-Rest was contained within the novel arm (Fig. 3c, top), a place the animal had not yet visited (that is, trajectory preplay). Moreover, in 79.8% of the trajectory preplay cases the shuffling procedures resulted in lines that were significantly less or not at all contained within the novel arm (that is, not preplay; Supplementary Information). The remaining trajectories decoded during Fam-Rest represented replay of the familiar track (64.15%; Fig. 3c, middle) or spanned the joint familiar track/novel arm space (19.49%; Fig. 3c, bottom). Means of absolute rank-order correlations between spiking activity and novel arm templates (Fig. 2a) restricted during epochs of trajectory preplay were significantly larger than those between spiking activity and familiar track templates calculated during the same epochs (0.75 versus 0.59, $P < 10^{-4}$). Overall, these results support the existence of the preplay phenomenon.

To investigate the possibility that preplay of novel arm place cell sequences during Fam-Rest depends on the prior run experience on the familiar track, mice with no prior experience on any linear track were placed in a high-walled sleep box and recorded while resting/sleeping. The animals were then transferred to a novel isolated linear track that was in the same room but could not be seen from inside the box, and the recording continued during *de novo* formation of place cells (Supplementary Fig. 2b, *de novo* session). We found that in a relatively large proportion (16.1%) of spiking events identified during sleep/rest in the sleep box, the neuronal firing sequences were significantly correlated with the place cell sequences observed during the first run session on the novel track (Fig. 4A, B and Methods); this was the case for all four individual mice (Supplementary Fig. 7). Preplay events were associated with the ripple occurrence (Fig. 4C). The place cells established on the novel track in the *de novo* session were more dynamic (median $r = 0.42$; Fig. 4D, blue) than in Contig-Run (median $r = 0.62$, $P < 0.016$; Fig. 2e, right, blue).

We have demonstrated that a significant number of temporal firing sequences of CA1 cells during resting periods of a familiar track exploration that preceded a novel track exploration in the same general environment were correlated with the place cell sequences of the novel track rather than the familiar track. This phenomenon, preplay, is temporally opposite to the process of replay[3–10,19,20], when activity during rest or sleep periods recapitulates place cell sequences that have already occurred during previous explorations. Preplay differs fundamentally from replay because it occurs before exploration of novel tracks.

Although our recordings were carried out in CA1, we believe that what we observed could be a reflection of the output of the recurrent cellular assemblies from upstream regions (CA3 or entorhinal cortex).

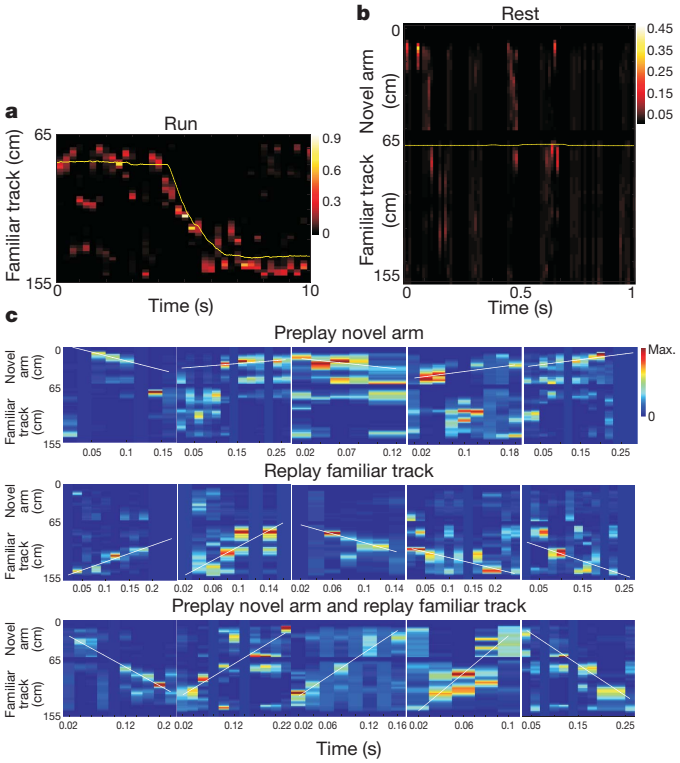


Figure 3 | Bayesian reconstruction of the animal's trajectory in the familiar track (replay) and novel arm (preplay). **a**, Position reconstruction of a one-lap run on the familiar track from the ensemble place cell activity during Fam-Run. The heat map displays the reconstructed position of the animal using ensemble place cell activity during the run (250-ms bins; animal velocity, >5 cm s^{-1}). The yellow line indicates the actual trajectory of the animal during Fam-Run. **b**, Example of virtual trajectory reconstruction (familiar track and novel arm) from the ensemble place cell activity during Fam-Rest at the ends of the familiar track (20-ms bins; animal velocity, <5 cm s^{-1}) before barrier removal and novel arm exploration. The yellow line reflects the spatial location of the animal in time: the animal was immobile at the junction end of the familiar track. The time-compressed ($\sim$5 m s^{-1}) trajectory reconstruction often 'jumps' over the barrier (top of the figure) into the novel arm area. At around 0.5 s, a preplay of the novel arm initiated from the distal (free) end of the novel arm 'propagates' towards the location of the animal. **c**, Examples of preplay of the novel arm (top), replay of the familiar track (middle) and preplay of the novel arm together with replay of the familiar track (bottom) during Fam-Rest. All conditions are the same as in **b**. The white line shows the linear fit maximizing the likelihood along the virtual trajectory. Colour bars indicate probability of trajectory reconstruction.

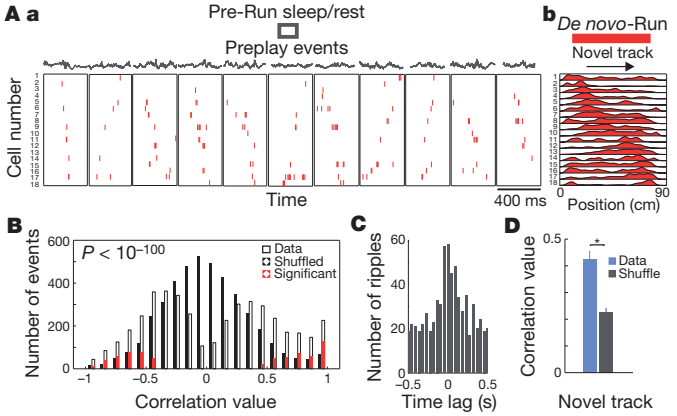


Figure 4 | Preplay of novel place cell sequences before any linear track experience. **A**, Sleep/rest session in the sleep box (Pre-Run sleep/rest) before the first run session on a linear track (*De novo*-Run). Display format is the same as in Fig. 1. **A**, **a**, Representative spiking events in the forward or reverse order during Pre-Run sleep/rest in 400-ms time windows. **A**, **b**, Place cell sequences on the novel track (red) during the *De novo*-Run session. Each row represents one cell in which the activity was normalized to the maximum firing rate. One run direction in one animal is shown. The median number of place cells active on the novel track participating in preplay events is six. **B**, Distribution of spiking events in Pre-Run sleep/rest as a function of the rank-order correlation with the place cell sequence template of the novel track. Display format is the same as in Fig. 2a. **C**, Cross-correlation between preplay events and ripple epochs during Pre-Run sleep/rest. **D**, Stability of place cell spatial tuning across the novel track experience. Display format is the same as in Fig. 2e (novel arm). Error bars, s.e.m.; asterisk indicates significant difference.

During running on a familiar track, some of the cells in the postulated upstream cellular assemblies fire sequentially at spatial locations while others, although connected anatomically to these cells, remain silent. The lack of expression of preplay sequences during Fam-Run may reflect their state-dependent suppression or subthreshold activation during these exploratory behaviours. Owing to increased net excitation during rest periods predominantly during ripples[21], some of these silent cells together with some of the familiar track cells are activated above threshold and fire in a certain sequence. Their sequence of activation may be determined in part by their functional connectivity within the hippocampal formation network. Some of these sequences may in turn be activated on a novel track as place cell sequences (Supplementary Fig. 8). The activation of the novel place cell sequences during running may strengthen their pre-existing assembly organization manifested during preplay.

It could be argued that during Contig-Run the animals simply considered the novel arm to be an extension of the familiar arm and, thus, what we considered to be preplay events were replays of the previous runs on the familiar track. If this was the case, preplay events would not be expected to be found when the experience of the familiar track run is eliminated. This idea was refuted by the demonstration of frequent preplay events in the sleep box before the mice were transferred onto a novel linear track (*de novo* condition). Under this condition, the place cell sequences were more dynamic and a higher proportion of all spiking events correlated with the place cell sequences in these runs than in the later runs on novel linear tracks. These results suggest a shift in the relative contribution of internal[22,23] and external drives in the formation of place cell sequences on encounter with a novel track. In the early phase, internal drives originating in the dynamic cellular assembly activities, which probably reflect numerous past experiences distinct from the current one and expressed as preplay, may have a greater role, whereas in the late phase, external drives that come from the specific set of stimuli of the current experience may dominate. Thus, place cell sequences on novel tracks seem to be products of a dynamic interplay between the internal and external drives.

Several previous studies did not reveal preplay[7,8,10,20]. Although it is difficult to pinpoint the apparent discrepancies between these studies and the present one, we suggest that the use of insufficiently sensitive methods (pairwise correlations) by some studies[7,8,20] and small sample sizes by others[10] might have precluded detection of preplay in previous work (see Supplementary Information for details). Data from the *de novo* condition (Fig. 4), in which we observed an even higher proportion of preplay events, have not been reported previously.

Our data showed that novel preplay events coexist in disjunction with familiar replay events during the rest periods on the familiar track. This and the finding that these preplay and replay events together make up fewer than one-quarter of all detected spiking events suggest that they are part of a dynamic repertoire of temporal sequences in the hippocampus that are past-experience dependent (replay) or future-experience expectant[24] (preplay). Post-experience replay of place cell sequences during resting[3–6] or slow-wave sleep[8–10] has been proposed to have an important role in memory consolidation[11,12]. The temporal preplay of new place cell sequences during resting or sleep is consistent with a predictive function for the hippocampal formation[25] and may contribute to accelerating learning[26]

when a new experience is introduced in multiple steps of increasing novelty.

METHODS SUMMARY

We recorded place cells from the CA1 area of the hippocampus with six independently movable tetrodes in four mice during sleep/rest sessions in the sleep box before any experience on linear tracks and during the first run session on a novel track. Following familiarization with the linear track, animals were subsequently allowed to explore a continuous (L-shaped) track in which the now familiar track and a new novel arm were made contiguous. To quantify the significance of the preplay and replay processes, spiking events in which at least four cells were active were detected during sleep/rest (speed, $<1\ \mathrm{cm\ s^{-1}}$) periods in the sleep box or awake rest (speed, $<2\ \mathrm{cm\ s^{-1}}$) periods at the ends of the familiar track and novel arm, predominantly during ripple epochs.

We calculated statistical significance at the $P<0.025$ level for each event by comparing the rank-order correlation between the event sequence and the place cell sequence (template) with the distribution of correlation values from 200 templates obtained by shuffling the original order of the place cells. Proportions of significant events were calculated as the ratio between the number of significant events and the total number of spiking events. We calculated the overall significance of preplay or replay processes by comparing the distribution of correlation values of all events with the distribution of correlation values of shuffled templates (Kolmogorov–Smirnov test). The significance of the proportion of significant events out of the total number of spiking events was determined as the binomial probability of observing the number of significant events (as successes) from the total number of spiking events (as independent trials), with a probability of success of 0.025 in any given trial. We reconstructed the position of the animal from the spiking activity emitted during resting periods using Bayesian decoding procedures[6].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 4 December 2009; accepted 29 October 2010.
Published online 22 December 2010.

1. O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map* (Oxford Univ. Press, 1978).
2. Wilson, M. A. & McNaughton, B. L. Dynamics of the hippocampal ensemble code for space. *Science* **261,** 1055–1058 (1993).
3. Foster, D. J. & Wilson, M. A. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* **440,** 680–683 (2006).
4. Diba, K. & Buzsaki, G. Forward and reverse hippocampal place-cell sequences during ripples. *Nature Neurosci.* **10,** 1241–1242 (2007).
5. Karlsson, M. P. & Frank, L. M. Awake replay of remote experiences in the hippocampus. *Nature Neurosci.* **12,** 913–918 (2009).
6. Davidson, T. J., Kloosterman, F. & Wilson, M. A. Hippocampal replay of extended experience. *Neuron* **63,** 497–507 (2009).
7. Wilson, M. A. & McNaughton, B. L. Reactivation of hippocampal ensemble memories during sleep. *Science* **265,** 676–679 (1994).
8. Skaggs, W. E. & McNaughton, B. L. Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science* **271,** 1870–1873 (1996).
9. Nádasdy, Z., Hirase, H., Czurko, A., Csicsvari, J. & Buzsaki, G. Replay and time compression of recurring spike sequences in the hippocampus. *J. Neurosci.* **19,** 9497–9507 (1999).
10. Lee, A. K. & Wilson, M. A. Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron* **36,** 1183–1194 (2002).
11. Buzsáki, G. Two-stage model of memory trace formation: a role for "noisy" brain states. *Neuroscience* **31,** 551–570 (1989).
12. Nakashiba, T., Buhl, D. L., McHugh, T. J. & Tonegawa, S. Hippocampal CA3 output is crucial for ripple-associated reactivation and consolidation of memory. *Neuron* **62,** 781–787 (2009).
13. Hebb, D. O. *The Organization of Behavior: A Neuropsychological Theory* (Wiley, 1949).
14. Harris, K. D., Csicsvari, J., Hirase, H., Dragoi, G. & Buzsaki, G. Organization of cell assemblies in the hippocampus. *Nature* **424,** 552–556 (2003).
15. Dragoi, G. & Buzsaki, G. Temporal encoding of place sequences by hippocampal cell assemblies. *Neuron* **50,** 145–157 (2006).
16. Thompson, L. T. & Best, P. J. Place cells and silent cells in the hippocampus of freely-behaving rats. *J. Neurosci.* **9,** 2382–2390 (1989).
17. O'Neill, J., Senior, T. & Csicsvari, J. Place-selective firing of CA1 pyramidal cells during sharp wave/ripple network patterns in exploratory behavior. *Neuron* **49,** 143–155 (2006).
18. Zhang, K., Ginzburg, I., McNaughton, B. L. & Sejnowski, T. J. Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. *J. Neurophysiol.* **79,** 1017–1044 (1998).
19. Johnson, A. & Redish, A. D. Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *J. Neurosci.* **27,** 12176–12189 (2007).
20. Kudrimoti, H. S., Barnes, C. A. & McNaughton, B. L. Reactivation of hippocampal cell assemblies: effects of behavioral state, experience, and EEG dynamics. *J. Neurosci.* **19,** 4090–4101 (1999).
21. Csicsvari, J., Hirase, H., Czurko, A., Mamiya, A. & Buzsaki, G. Oscillatory coupling of hippocampal pyramidal cells and interneurons in the behaving rat. *J. Neurosci.* **19,** 274–287 (1999).
22. Dragoi, G., Harris, K. D. & Buzsaki, G. Place representation within hippocampal networks is modified by long-term potentiation. *Neuron* **39,** 843–853 (2003).
23. Pastalkova, E., Itskov, V., Amarasingham, A. & Buzsaki, G. Internally generated cell assembly sequences in the rat hippocampus. *Science* **321,** 1322–1327 (2008).
24. Black, J. E. & Greenough, W. T. *Advances in Developmental Psychology* (Lawrence Erlbaum, 1986).
25. Hassabis, D., Kumaran, D., Vann, S. D. & Maguire, E. A. Patients with hippocampal amnesia cannot imagine new experiences. *Proc. Natl Acad. Sci. USA* **104,** 1726–1731 (2007).
26. Tse, D. *et al.* Schemas and memory consolidation. *Science* **316,** 76–82 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. A. Wilson for assistance with data acquisition, discussions and comments on an earlier version of the manuscript; J. O'Keefe, A. Siapas, F. Kloosterman, D. L. Buhl for comments on earlier versions of the manuscript; and F. Kloosterman for providing assistance with the line detection for the Bayesian decoding. This work was supported by NIH grants R01-MH078821 and P50-MH58880 to S.T., who was an HHMI Investigator in an earlier part of this study.

Author Contributions S.T. and G.D. conceived the project jointly. G.D. designed and performed the experiments and the analyses. G.D. and S.T. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to G.D. (gdragoi@mit.edu) or S.T. (tonegawa@mit.edu).

METHODS

Surgery and experimental design. Electrophysiological recordings were performed on four C57BL/6 mice (strain NR1-floxed[27]) with ages between 18 and 22 weeks. All animals were implanted under Avertin anaesthesia with six independently movable tetrodes aiming for the CA1 area of the right hippocampus (1.5–2 mm posterior to bregma and 1–2 mm lateral to the midline; Supplementary Fig. 1). The reference electrode was implanted posterior to lambda over the cerebellum. During the following week of recovery, the electrodes were advanced daily while animals rested in a small, walled sleeping box ($12 \times 20\,cm^2$, 35 cm high). The animal position was monitored by means of two infrared diodes attached to the headstage.

The experimental apparatus consisted of a $90 \times 65\,cm^2$ rectangular, walled, linear track maze. All tracks were 4 cm wide at the bottom and 8–9 cm wide at the top, and all linear track walls were 10 cm high. Experimental sessions were conducted while the animals explored for chocolate sprinkle rewards placed always at the ends of the corresponding linear tracks (one sprinkle at each end of the track on each lap). Neuronal activity was recorded in naive animals (four mice) during the sleep/rest session in the sleep box immediately preceding the first experience on linear tracks, and continued (Fig. 4) during the first run session on a novel track. After familiarization with the linear track, the animals went through a recording session of 15–60 min (Fam session), and the recordings continued for the next 34–42 min (Contig session) while the animals explored an L-shaped track for the first time. In this track, the familiar arm and the novel arm were made contiguous by removing the barrier that had separated them (Fig. 1). For the purpose of analysing the recording data, the Fam session was further divided into Fam-Run, in which the animals ran through the track (velocity of animal's movement was higher than $5\,cm\,s^{-1}$), and Fam-Rest, where the animals took awake rests at the ends of the track (velocity of animal's movement was less than $2\,cm\,s^{-1}$). During resting periods, the animals consumed the chocolate sprinkle and groomed, but mostly they were still until they self-initiated the next lap of run on the linear track. After completion of the experiments, the brains of all mice were perfused, fixed, sectioned and stained using nuclear fast red (Supplementary Fig. 1) or cresyl violet for electrode track reconstruction.

Recordings and single-unit analysis. A total of 87 neurons were recorded from the CA1 area of the hippocampus in four mice during the Fam and Contig sessions (Supplementary Tables 1–3). A total of 69 CA1 neurons were recorded from the four mice in the *de novo* condition (26, 20, 10 and 13 cells, respectively). Single cells were identified and isolated using the manual clustering method Xclust[2] and the application of cluster quality measurements[28]. Pyramidal cells were distinguished from interneurons on the basis of spike width, average rate and autocorrelations[22].

Place fields were computed as the ratio between the number of spikes and the time spent in 2-cm bins along the track, smoothed with a Gaussian kernel with a standard deviation of 2 cm. Bins where the animal spent a total of less than 0.1 s and periods during which the animal's velocity was below $5\,cm\,s^{-1}$ were excluded. Place field length and peak rate were calculated after separating the direction of movement and linearizing the trajectory of the animal. Linearized place fields were defined as areas with a localized increase in firing rate above 1 Hz for at least five contiguous bins (10 cm). The place field peak rate and location were given by the rate and location of the bin with the highest ratio between spike counts and time spent. Place field borders were defined as the points where the firing rate became less than 10% of the peak firing rate or 1 Hz (whichever was bigger) for at least 2 cm.

Local field potential analysis. Ripple oscillations were detected during sleep/rest periods in the sleep box and during rest periods at the ends of the tracks. The electroencephalography signal was filtered (120–200 Hz) and ripple-band amplitude was computed using the Hilbert transform. Ripple epochs with maximal amplitude more than 5 s.d. above the mean, beginning and ending at 1 s.d. were detected. The time of ripple occurrence (Figs 2c and 4C) was the time of its maximal amplitude. The proportion of ripples with which cells with place fields on the novel arm of the L-shaped track fired in the preceding session (Supplementary Fig. 3) was calculated for each qualifying cell as the ratio between the number of ripples during which the cell fired at least one spike and the total number of ripples during the corresponding exploratory session.

Preplay and replay analyses. To analyse the preplay and replay processes, spiking events were detected during Pre-Run sleep/rest periods in the sleep box (*de novo* condition; velocity, $<1\,cm\,s^{-1}$) or during awake rest periods at the ends of the running tracks (Contig condition; velocity, $<2\,cm\,s^{-1}$). A spiking event was defined as a transient increase in the firing activity of a population of at least four different place cells within a temporal window preceded and followed by at least 50 ms of silence. Overall, similar results were obtained using 50-, 60-, 75- and 100-ms time windows. The spikes of all the place cells active on the novel track that were emitted during the Pre-Run sleep/rest in the box for the *de novo* condition as well as the spikes of all the place cells active on the familiar track or the novel arm that were emitted during Fam-Rest session at the two ends of the familiar track for Contig condition were respectively sorted by time and further used for the detection of the spiking events.

All four animals exhibited a significant number of spiking events in the Pre-Run session of the *de novo* condition. Three of the four animals (mice 1–3) exhibited a significant number of spiking events in the Contig condition, the remaining animal (mouse 4) having a below-threshold number of simultaneously active CA1 place cells. The time of the spiking event used to compute the cross-correlation with ripple epoch occurrence (Figs 2c and 4C) was the average time of all spikes comprising the spiking event. The place cell sequences (templates) were calculated for each direction of the animal's movement and for each run session (*De novo*-Run, Fam-Run and Contig-Run) by ordering the spatial location of the place field peaks that were above 1 Hz. For place cells with multiple place fields above 1 Hz on a particular arm or track in the Contig condition (six of 52 place cells active on the novel arm in the two directions, or 12%: two for each direction in mouse 1, one in mouse 2 and one in mouse 3), only the place field corresponding to the peak firing rate of the place cell on that arm or track was considered for the construction of the template of that particular arm or track, to be consistent with all the previous studies that used spatial templates to demonstrate replay during sleep or awake rest[3,4,10]. Place cells with fields on both the novel arm in the Contig-Run session and the familiar track in the Fam-Run session participated in the construction of both the novel arm and familiar track templates.

Statistical significance was calculated for each event by comparing the rank-order correlation between the sequence of cells' firing during the event (that is, event sequence) and the place cell sequence (template), on the one hand, and the distribution of correlation values between the event sequence and 200 surrogate templates obtained by shuffling the order of place cells, on the other[4] (Fig. 2a). The significance level was set at 0.025 to control for multiple comparisons (two directions of run). The proportions of significant events (preplay novel track, preplay novel arm (Fig. 2b), replay novel arm and replay familiar track) were each calculated as the ratio between the number of significant events and the total number of spiking events in which at least four corresponding place cells were active[4]. Corresponding familiar track templates (Fig. 2h) were constructed by ordering the location of peak firing on the familiar track during Fam-Run (no minimum threshold of firing) of all place cells that subsequently fired on the novel arm. Cells comprising the corresponding familiar track templates are the same as those comprising the novel arm templates. We note that these corresponding familiar track templates are different from the ones used in Figs 1 and 2a–g, which were constructed by ordering the peak firing of all place cells active on the familiar track >1 Hz.

The overall significance of the preplay (Fig. 2a) or replay process was calculated by comparing the distribution of correlation values of all events relative to the original template with the distribution of correlation values relative to the shuffled surrogate templates, using the Kolmogorov–Smirnov test[3]. Quantification of the replay versus preplay events during the Fam-Run session (Fig. 2f, g) was performed as described above using different spatial templates for the familiar track and the novel arm. All spiking events were correlated with both the novel arm and the familiar track templates. Events significantly correlated only with familiar track or with novel arm templates were considered pure replay and pure preplay, respectively. The template specificity index was calculated for each event as the difference between the absolute value of the event's correlation with the novel arm template (preplay, high positive index) and the event's correlation with the familiar track template (replay, high negative index). For the purpose of displaying the template specificity index, events correlated with the novel arm but not with the familiar track templates were considered preplay and events correlated with the familiar track but not with the novel arm templates were considered replay (Fig. 2g). Additionally, events correlated with both the familiar track and the novel arm templates formed a third group, preplay/replay events, displayed in yellow in the inset of Fig. 2g.

Correlations between pairs of familiar track and novel arm templates (Fig. 2h) were performed using modified familiar track templates that were constructed using the location of peak firing (>0 Hz) of only those cells that had place fields on the novel arm (peak rate, >1 Hz). The lack of significant correlation in this case demonstrates that the novel arm place cell sequence is not simply a transposition of a familiar track place cell sequence on the novel arm.

We also identified neurons that did not fire during Fam-Run, that activated during Fam-Rest events and that corresponded to trajectories on the novel arm during Contig-Run (silent cells). We calculated the correlation between the order in which they fired during Fam-Rest events and their spatial sequence as new place cells on the novel arm during Contig-Run, as previously explained. Owing to the low absolute number of silent neurons, only triplets of cells were available for further analysis ($n = 24$). The proportion of events perfectly matching the spatial template was compared with the proportion of by-chance perfect matching (0.33).

Stability of place cell maps. Stabilities of place cell firing on the familiar track before and after barrier removal as well as on the novel track (*de novo* condition) and the novel arm (Contig condition) in the beginning versus the end of the run session were assessed by calculating, for each place cell and each direction, a correlation between the spatial firing in the corresponding paired situations (before versus after barrier removal for the familiar track or the first four laps versus the last four laps of the *De novo*-Run or Contig-Run session for the novel track or arm, respectively). The place cell activity was not partitioned in place fields; rather, the whole activity on the particular track or arm was considered separately for each cell and direction (average correlations are shown in Figs 2e and 4D, blue bars). In addition, we calculated the same type of correlation after shuffling the identity of the cell in one member of the correlation (once for each different cell; average correlations are in Figs 2e and 4D, black bars). Shuffle results (Figs 2e and 4D, black bars) were computed as correlation between spatial tuning of cells on the familiar track during Fam-Run and spatial tuning of all other simultaneously recorded cells on the familiar arm during Contig-Run (familiar track group; Fig. 2e, left), or correlation between spatial tuning of cells on the novel arm (or novel track) during the beginning of Contig-Run (or *De novo*-Run) and spatial tuning of all the other simultaneously recorded cells on the novel arm (or novel track) during the end of Contig-Run (novel arm group; Fig. 2e, right) or *De novo*-Run (Fig. 4D). Original and shuffled correlations were compared using the rank-sum test. The average number of laps (traversal of the novel track in both directions) per session was 20.5 in *De novo*-Run (21, 16, 27 and 18 in the four mice) and 16.3 in Contig-Run (13, 14 and 22 in the three mice).

Bayesian reconstruction of actual and virtual trajectories. For each cell, we calculated a linearized spatial tuning curve on the familiar track during the Fam-Run session and a linearized spatial tuning curve on the novel arm during the Contig-Run session. The tuning curves were constructed in 2-cm bins from spikes emitted in both run directions at velocities higher than 5 cm s^{-1}, and were smoothed with a Gaussian kernel with a standard deviation of 2 cm. We constructed a joint spatial tuning curve for each cell by juxtaposing the spatial tuning curve on the familiar track during the Fam-Run session and the spatial tuning curve on the novel arm during the Contig-Run session. We also detected for each cell all the spiking activity emitted at velocities below 5 cm s^{-1} during the Fam-Rest session, where replay and preplay events where shown to occur using the rank-order correlation method. We used a Bayesian reconstruction algorithm[6,18] to decode the virtual position of the animal from the spiking activity during Fam-Rest (Fig. 3b) in non-overlapping, 20-ms bins using the joint spatial tuning curves. We then extracted epochs of reconstructed trajectory matching the time of the spiking events as detected using multiunit activity of place cells from the familiar track and novel arm (rank-order correlation method; see 'Preplay and replay analyses', above).

We used two shuffling procedures to measure the quality of the Bayesian decoding. In the first shuffling procedure, for each event, the original time-bin columns of the probability distribution function (PDF) were replaced with an equal number of time-bin columns randomly extracted from a pool containing the time-bin columns of all PDFs of all detected events[6]. The shuffling procedure was repeated 500 times. In the second shuffling procedure, the identity of the place cells was randomly shuffled 100 times and new PDFs were calculated for all events. For all original and shuffled PDFs, a line was fitted to the data using a previously described line-finding algorithm[6]. Lines fitted to the original and shuffled data were compared using slope, spatial extent, location on the track and probability score. We defined replay and preplay as the epochs of Fam-Rest in which the reconstructed trajectory was located on the familiar track or the novel arm, respectively. The trajectory was defined across a set of position estimates during the corresponding epoch (Fig. 3c). Only epochs that lasted at least 60 ms (three bins) and which contained reconstructed trajectories spanning at least 10 cm were considered for further analysis. Trajectories for which 75% or more of their length was located on the familiar track were considered to represent replay of an animal's trajectory on the familiar track (Fig. 3c, middle), and trajectories for which 75% or more of their length was located in the novel arm were considered to represent preplay of the animal's future trajectory on the novel arm (Fig. 3c, top). The remaining events were considered preplay–replay (Fig. 3c, bottom).

An epoch was considered significant if the new line was less than 75% contained in the familiar track for replay or novel arm for preplay in at least 95% of the shuffled cases. For each epoch that was significant for replay or preplay using the reconstruction method, we retrieved the value of the rank-order correlation between the neuronal firing sequences and the familiar track and novel arm spatial templates as calculated using the rank-order correlation method. We compared the absolute correlation values between the epoch's firing sequences and familiar track templates with the absolute correlation values between the same epoch's firing sequences and novel arm templates. We also reconstructed the trajectory of the animal on the familiar track from the spiking activity during the Fam-Run session at velocities above 5 cm s^{-1} in 250-ms bins using the spatial tuning curves on the familiar track[6,18] (Fig. 3a) to validate the decoding procedure.

27. Tsien, J. Z., Huerta, P. T. & Tonegawa, S. The essential role of hippocampal CA1 NMDA receptor-dependent synaptic plasticity in spatial memory. *Cell* **87,** 1327–1338 (1996).
28. Schmitzer-Torbert, N., Jackson, J., Henze, D., Harris, K. & Redish, A. D. Quantitative measures of cluster quality for use in extracellular recordings. *Neuroscience* **131,** 1–11 (2005).

LETTER

doi:10.1038/nature09600

Development of asymmetric inhibition underlying direction selectivity in the retina

Wei Wei[1], Aaron M. Hamby[1], Kaili Zhou[1] & Marla B. Feller[1,2]

Establishing precise synaptic connections is crucial to the development of functional neural circuits. The direction-selective circuit in the retina relies upon highly selective wiring of inhibitory inputs from starburst amacrine cells[1] (SACs) onto four subtypes of ON–OFF direction-selective ganglion cells (DSGCs), each preferring motion in one of four cardinal directions[2]. It has been reported in rabbit that the SACs on the 'null' sides of DSGCs form functional GABA (γ-aminobutyric acid)-mediated synapses, whereas those on the preferred sides do not[3]. However, it is not known how the asymmetric wiring between SACs and DSGCs is established during development. Here we report that in transgenic mice with cell-type-specific labelling, the synaptic connections from SACs to DSGCs were of equal strength during the first postnatal week, regardless of whether the SAC was located on the preferred or null side of the DSGC. However, by the end of the second postnatal week, the strength of the synapses made from SACs on the null side of a DSGC significantly increased whereas those made from SACs located on the preferred side remained constant. Blocking retinal activity by intraocular injections of muscimol or gabazine during this period did not alter the development of direction selectivity. Hence, the asymmetric inhibition between the SACs and DSGCs is achieved by a developmental program that specifically strengthens the GABA-mediated inputs from SACs located on the null side, in a manner not dependent on neural activity.

The ability to detect motion in the visual scene is a fundamental computation in the visual system that is first performed in the retina. Motion direction is encoded by DSGCs, which fire a maximum number of action potentials during movement in their preferred direction, but fire minimally for movement in the opposite, or null, direction[4,5]. In the mammalian retina, the directional preference of an ON–OFF DSGC is caused by asymmetric inhibitory inputs: movement in the null direction causes strong inhibition that effectively shunts light-evoked excitatory inputs. Indeed, blocking $GABA_A$ receptors abolishes the directionality of DSGCs by increasing spiking in response to null-direction motion[6–8]. Null-side inhibition is thought to arise from SACs because their processes cofasciculate with DSGC dendrites[9,10], where they form direct GABAergic synapses[3], and because ablation of SACs eliminates the directional preference of DSGCs[11,12].

How SAC–DSGC synapses are organized to provide asymmetric inhibition has been an intriguing but difficult question because no apparent asymmetry is detected in the morphology or the distribution of synaptic markers in DSGCs and SACs[13–15]. The first and only piece of evidence for the synaptic basis of asymmetric inhibition came from a functional study between SAC and DSGC pairs in rabbit retina[3], which suggested that SACs on the null side provide inhibitory inputs to the DSGCs but that those on the preferred side do not. Whether this asymmetric inhibition exists in the mouse is not known. In addition, because the directional preference of an ON–OFF DSGC is present by eye opening[16–18] and the identification of DSGCs and their preferred directions is almost impossible before the onset of the light response, little is known about the developmental program that shapes the SAC–DSGC synapses.

Here we use paired recordings and morphological reconstructions from a double-transgenic mouse line that selectively expresses two variants of green fluorescent protein (GFP) in SACs and nasal-preferring ON–OFF DSGCs (nDSGCs) to characterize the organization and the development of the precise wiring between SACs and DSGCs. These mice were generated by crossing two existing lines: *Drd4–GFP* mice, where *Drd4* promoter-driven GFP expression is restricted to nDSGCs[19], and *mGluR2–GFP* mice (*mGluR2* also known as *Grm2*), where a membrane-tethered human interleukin-2α/GFP fusion protein is expressed specifically in SACs in the retina (Fig. 1a)[20].

To detect functional GABAergic synapses between SACs and DSGCs, we performed targeted whole-cell voltage-clamp recordings from SAC–DSGC pairs in whole-mount retinas. To isolate GABAergic synapses, paired recordings were carried out in the presence of drugs that block excitatory synaptic transmission (Fig. 1b). Alexa dyes were included in the recording pipettes to visualize the dendritic morphology of the recorded pairs (Fig. 1c). Only pairs with overlapping dendritic fields were used for analysis.

Paired recordings were carried out in postnatal-day-4 (P4), P7, P14 and adult mice. At P4, GABAergic currents elicited by SAC depolarization were detected in nDSGCs in 64% of pairs (16 of 25 pairs; Fig. 1b, d), indicating that synapse formation between SACs and nDSGCs occurred before and during the first postnatal week, confirming previous findings[10]. By P7, nearly all pairs showed unitary GABAergic connections (P7: 85%, 29 of 34 pairs), and this high level of connectivity persisted into adulthood (P14–48: 91%, 41 of 45 pairs; Fig. 1d). The evoked response was completely blocked by the $GABA_A$ receptor antagonist gabazine (5 μM, $n = 4$; data not shown), indicating that the GABAergic transmission between SACs and nDSGCs is mediated by $GABA_A$ receptors. We note that the finding that connections were readily detected between SACs located on the preferred side of DSGCs in adult mice is in contrast to previous findings in rabbit[3].

Though SACs located on both the preferred side and the null side formed GABAergic synapses with DSGCs, a significant asymmetry in the unitary synaptic strength emerged along the null-preferred axis during the second postnatal week. Synaptic strength was quantified as the $GABA_A$-receptor-mediated whole-cell conductance. These measurements were restricted to the null-side and preferred-side pairs that had similar amounts of overlap between SAC processes and DSGC dendrites. Unexpectedly, at P4 and P7 the GABAergic conductances from both groups were similar (Fig. 2). However, a significant increase in unitary conductance was detected in the null-side pairs but not in the preferred side pairs in retinas at P14 and older (Fig. 2). Hence, the establishment of the direction-selective circuits is mediated by an asymmetric increase in the strength of the unitary conductance between SACs and DSGCs in the week before eye opening.

The difference in GABAergic conductance from the null- and preferred-side SACs prompted us to examine two possibilities regarding the mechanisms underlying this strengthening. First we tested whether this functional asymmetry was correlated with the number or quality of contacts between SACs and nDSGCs, indicating a preferential adhesion

[1]Department of Molecular & Cell Biology, University of California, Berkeley, California 94720-3200, USA. [2]Helen Wills Neurosciences Institute, University of California, Berkeley, California 94720-3200, USA.

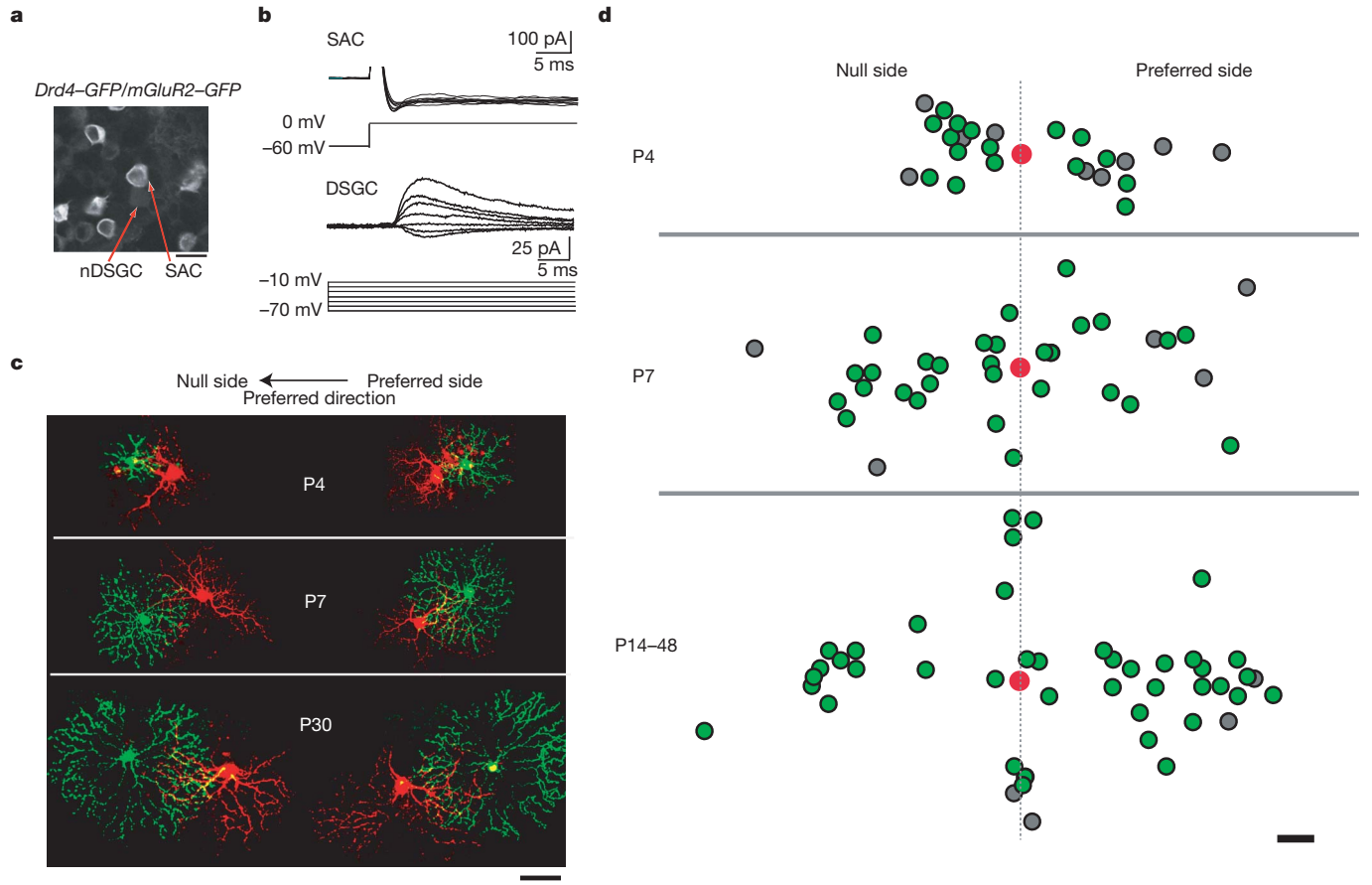


Figure 1 | nDSGCs receive direct GABAergic inputs from SACs located on the null and the preferred side from P4 until adult. **a**, Fluorescence image of the ganglion cell layer from a P30 *Drd4–GFP/mGluR2–GFP* mouse, showing the bright membrane-bound GFP expressed under the *mGluR2* promoter in the SACs and the dim cytoplasmic GFP driven by the *Drd4* promoter in the nDSGC. Scale bar, 25 μm. **b**, Paired whole-cell voltage-clamp recordings of GABAergic currents in a P4 nDSGC (lower traces) evoked by depolarization of a SAC from the null side (upper traces) in the presence of the NMDA (*N*-methyl-D-aspartate) receptor antagonist D(−)-2-amino-5-phosphonovaleric acid (AP5), α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)/kainate receptor antagonist 6,7-dinitroquinoxaline-2,3-dione (DNQX) and α4-containing nicotinic acetylcholine receptor antagonist dihydro-β-erythroidine (DHβE). SACs were depolarized from −60 to 0 mV, which reliably evoked an inward current in SACs. The postsynaptic GABAergic currents were recorded in DSGCs at different holding potentials to determine the current–voltage relationship of the conductance. **c**, Example images of synaptically connected, dye-filled SAC–DSGC pairs at P4, P7 and P30. The left-hand side shows pairs with SACs (green) located on the null side of the DSGCs (red). The right-hand side shows preferred-side pairs. Scale bar, 50 μm. **d**, Soma locations of the GABAergically connected SAC–nDSGC pairs along the null-preferred axis during development. Red spots represent the positions of DSGC cell bodies. The positions of SAC cell bodies that form GABAergic synapses with their respective nDSGCs are shown as green spots; the SAC cell bodies that were not connected to nDSGCs are shown as grey spots. All pairs had overlapping dendritic fields. Scale bar, 25 μm.

between SACs located on the null side and DSGCs[3], which is an important mechanism for dendritic differentiation and synaptogenesis in other systems[21,22]. After electrophysiological recording, the dendritic arborizations of the synaptically connected, Alexa-dye-filled SAC–nDSGC pairs from P14 to P48 were imaged live with a two-photon microscope and reconstructed using NEUROLUCIDA (Fig. 3a). We examined the overlapping region between the nDSGC dendrites and the distal portion (roughly the outer third) of the SAC processes enriched in varicosities, which are the sites of neurotransmitter release[9]. Crossing points between distal SAC processes and nDSGC dendrites were defined as 'contacts' (Fig. 3a, inset). A subset of contacts exhibited cofasciculation[9,10,23], which were defined as 2-μm segments along which the processes from the two cells remained in contact (Fig. 3a, inset). The null- and preferred-side SAC–nDSGC pairs showed a similar density of contacts (Fig. 3b) and cofasciculations (Fig. 3c). No asymmetry was found when all of the SAC processes were included in the above analysis (Supplementary Fig. 1). Therefore, the functional asymmetry in GABAergic synapses does not involve selective adhesion between null-side SAC processes and DSGC dendrites[24].

The second possibility we tested was whether spontaneous retinal activity during the second postnatal week has a role in the establishment of direction selectivity. DSGCs are depolarized by retinal waves, and activity could therefore potentially influence the synapse strengthening[25]. To this end, we first confirmed that the GFP-labelled nDSGCs in the *Drd4–GFP* mice showed a clear preference for nasal motion at eye opening (Fig. 4a) that was sensitive to the $GABA_A$ receptor antagonist gabazine (Supplementary Fig. 2), with a direction selectivity index similar to those recorded in the adult (Fig. 4b). We then injected muscimol, a $GABA_A$ receptor agonist, intravitreally into *Drd4–GFP* mice to block all spontaneous and evoked neural activity in the retina[26]. In the presence of muscimol, evoked synaptic transmission from SACs to nDSGCs and spontaneous activity in both cell types were completely suppressed (Fig. 4c and Supplementary Fig. 3a, b). The effectiveness of muscimol injection at blocking activity *in vivo* was confirmed by examining eye-specific segregation of retinogeniculate projections, which is an independent measure of retinal activity (Supplementary Fig. 3c, d), and the persistence of fluorescently labelled muscimol in the retina at 48 h post-injection (Supplementary Fig. 3e).

We assessed the responses of nDSGCs to stationary flashes and drifting gratings in P14–15 mice that had received repeated muscimol injections in the second postnatal week. Muscimol treatment did not prevent the development of direction-selective responses or significantly reduce

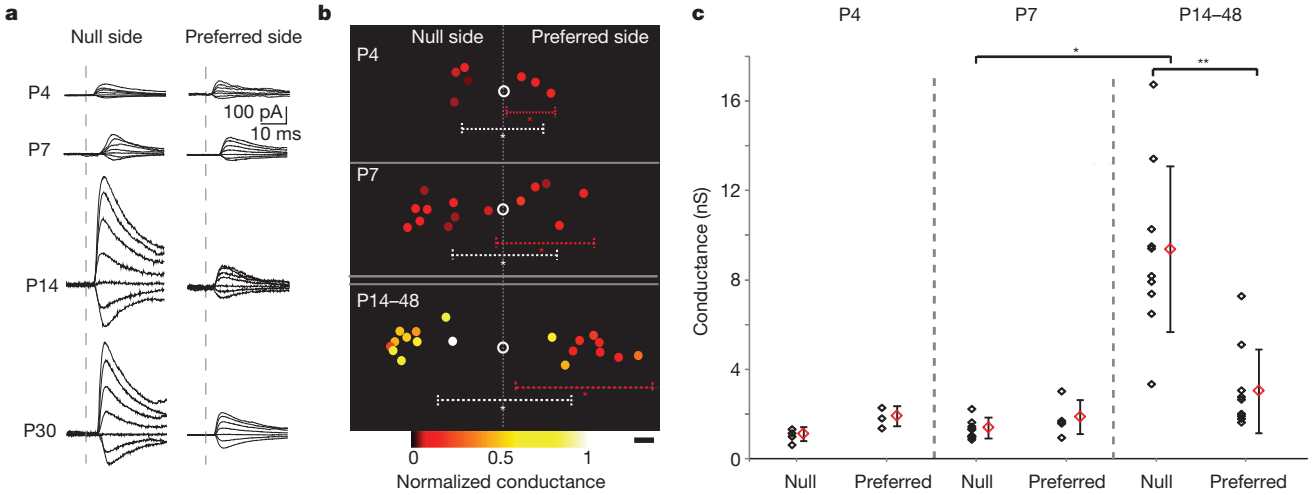


Figure 2 | GABAergic conductance in the null-side SAC–nDSGC pairs strengthens during the second postnatal week. **a**, Postsynaptic GABAergic currents in nDSGCs recorded at holding potentials between –70 and –10 mV in response to depolarization (as in Fig. 1b) of null-side (left) and preferred-side (right) SACs at P4, P7, P14 and P30. **b**, Relative soma positions of SAC–nDSGC pairs used for conductance analysis at P4, P7 and P14–48. Open circles represent nDSGC cell bodies. Filled circles are SAC somas colour-coded for conductance strength normalized to the maximum value across all ages. Dashed lines illustrate average dendritic arborization diameter, centred on the asterisks, for nDSGCs (white) and SACs (red; asterisks represent average soma locations). Scale bar, 25 µm. **c**, Summary plot of GABAergic conductances of the null- and preferred-side SAC–nDSGC pairs at P4, P7 and P14–48. Individual pairs (black) and mean ± s.d. (red) are shown. One-way analysis of variance: $P < 0.0001$; *t*-test: $^{*}P < 0.0001$, $^{**}P = 0.0003$.

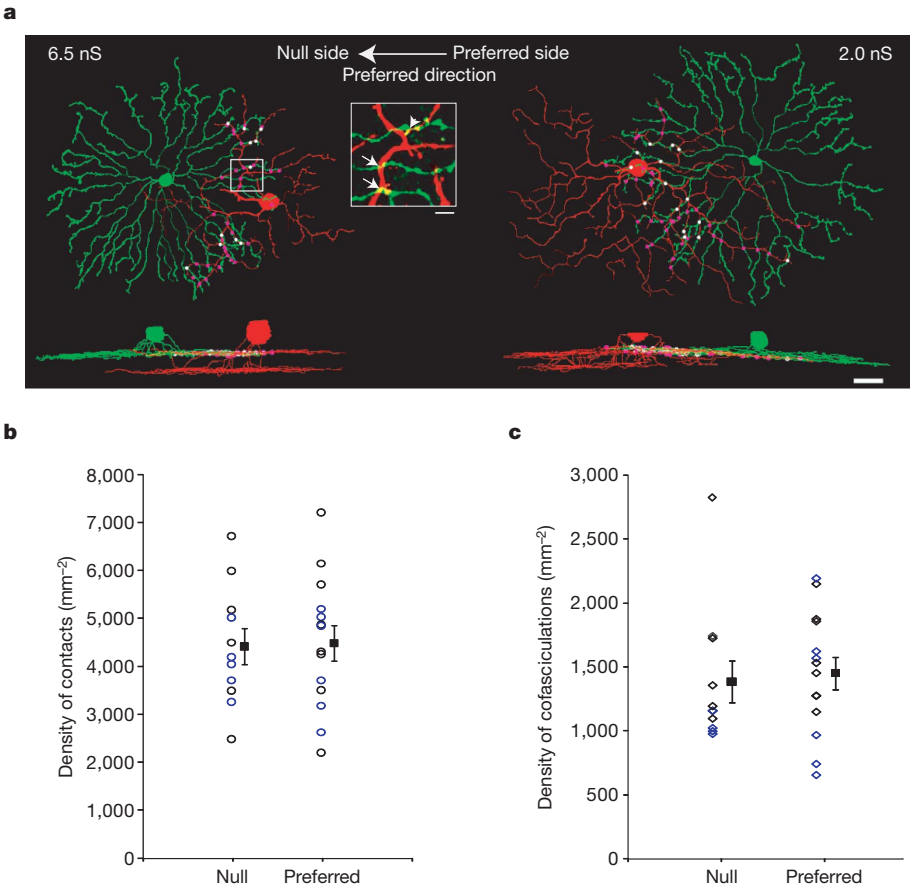


Figure 3 | Dendritic contacts and cofasciculations between SACs and nDSGCs occur at similar densities for the null- and preferred-side pairs. **a**, NEUROLUCIDA reconstructions of the dendrites from the on sublamina and side views of the complete dendritic arborizations from a null-side (left) and a preferred-side (right) pair of SACs and nDSGCs. Dots represent dendritic contacts, with cofasciculation segments coloured white and the rest coloured purple. The GABAergic conductances for the null- and preferred-side pairs are indicated. Scale bar, 25 µm. Inset, fluorescence image of the outlined region showing crossing contacts (arrows) and cofasciculation (arrowhead). Scale bar, 5 µm. **b**, Summary plot of the density of total contacts between DSGCs and distal SAC processes (roughly the outer third) from the null or preferred side from P14 to P48. Individual pairs and mean ± s.d. are shown. The data points for P28 and later are coloured blue, and the ones for before P28 are coloured black. **c**, Summary plot of the density of cofasciculations between nDSGCs and distal SAC processes from the same pairs as in **b**. Null- and preferred-side groups are not significantly different in **b** and **c**. $P > 0.7$, *t*-test.

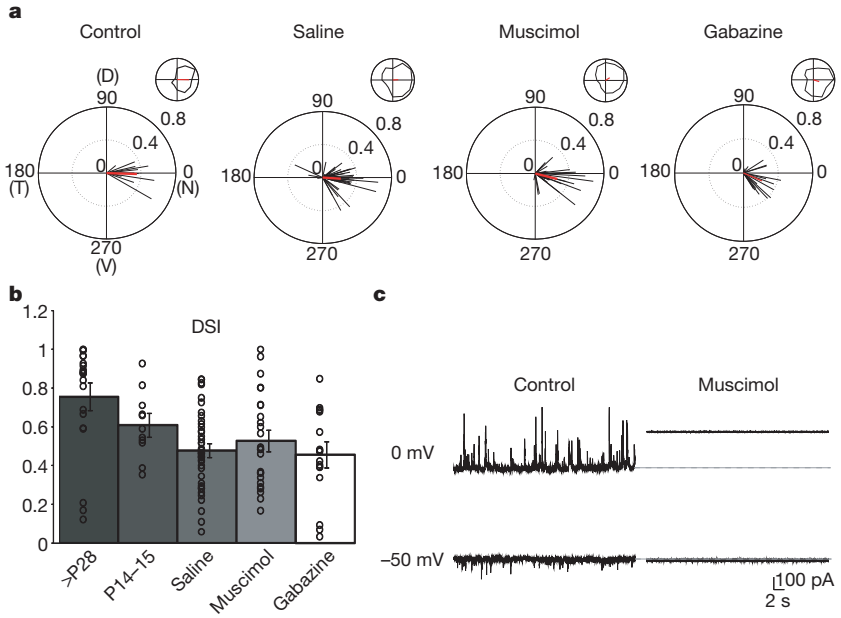


Figure 4 | Intraocular injections of muscimol or gabazine do not alter direction selectivity in nDSGCs. **a**, The normalized spike vector sums of nDSGCs in response to drifting gratings of 12 directions from P14–15 *Drd4–GFP* mice that received either no treatment (control) or intraocular injections of saline, muscimol or gabazine from P6 to P12. D, dorsal; N, nasal; T, temporal; V, ventral. The red lines are mean vector sums of all cells in each group. Insets, examples of normalized tuning curves of single cells, with corresponding vector sums represented as red lines of nDSGCs from each group. Control: $n = 4$ mice, 12 cells; saline: $n = 11$ mice, 43 cells; muscimol: $n = 12$ mice, 25 cells; gabazine: $n = 4$ mice, 17 cells. **b**, Summary plot of direction selectivity index (DSI) for adult (>P28), P14–15 untreated, saline, muscimol and gabazine-treated groups. Bars show mean ± s.e.; open circles represent individual cells. Adult data are reproduced from ref. 19. **c**, Example traces from whole-cell voltage-clamp recordings of inhibitory (upper traces, $V_H = 0$ mV) and excitatory (lower traces, $V_H = -50$ mV) currents from a P14 nDSGC in drug-free artificial cerebrospinal fluid (control, left) or artificial cerebrospinal fluid containing 100 μM muscimol (right). Deflections from baseline correspond to spontaneous synaptic currents. At depolarized potentials, application of muscimol activated a tonic current, which was measured as a change in the baseline holding current[26].

directional tuning of nDSGCs (Fig. 4a, b). Normal ON and OFF light responses were also present in the muscimol-treated group, although there was an increase in the number of cells that did not respond to gratings in the muscimol-treated group (Supplementary Fig. 4).

In visual cortex, activation of $GABA_A$ receptors is required for maturation of GABAergic synapses[27]. To test the hypothesis that $GABA_A$ receptor activation is required for the development of direction selectivity, we performed intravitreal injections of the $GABA_A$ receptor antagonist gabazine into *Drd4–GFP* mice during the second postnatal week. Gabazine treatment did not prevent the development of direction-selective responses of GFP-positive cells to drifting gratings (Fig. 4a, b). Therefore, the development of direction selectivity arises independently of the activation of $GABA_A$ receptors.

To begin exploring the synaptic basis of this increase in conductance between null-side SACs and DSGCs, we recorded the spontaneous inhibitory postsynaptic currents (IPSCs) from nDSGCs at P7 and P14. We found a significant increase in the frequency and no significant change in the amplitude of the GABAergic IPSCs, although there was a trend towards larger IPSC amplitudes at P14 (Supplementary Fig. 5). This result is consistent with the hypothesis that the stronger GABAergic unitary conductance for the null-side pairs is primarily due to increased numbers of functional GABAergic synapses. However, we cannot tell whether the spontaneous IPSCs originate from the null- or preferred-side SACs. Further study is required to determine the relative role increases in synapse number versus synapse strength have in the increase in the unitary conductance between null-side SACs and DSGCs.

Our study demonstrates that asymmetric inhibition arises during the second postnatal week through selective strengthening of the GABAergic conductance from SACs on the null sides of DSGCs. Morphological analysis revealed a similar degree of dendritic contact and cofasciculation between SACs on the null or preferred side, indicating that the synapse development is dissociated from physical encounters between SAC processes and DSGC dendrites, as was recently found in barrel cortex[28].

In addition, we found that blocking depolarization-induced activity or $GABA_A$ receptor activation did not affect the establishment of direction selectivity in the retina, in sharp contrast to direction-selective cells in the visual cortex[29]. This finding lends support to previous studies showing that early visual experience[16–18] or cholinergic retinal waves[18] are not involved in establishing retinal direction selectivity. Therefore, the mechanism underlying the development of retinal direction selectivity is an asymmetric increase in the strength of the inhibitory unitary conductance between SACs and DSGCs in the week before eye opening, without the establishment of asymmetrical dendritic contacts and independent of spontaneous neural activity.

METHODS SUMMARY

We performed dual whole-cell voltage-clamp recordings from SAC–nDSGC pairs in oxygenated artificial cerebrospinal fluid at 32–34 °C containing 119.0 mM NaCl, 26.2 mM $NaHCO_3$, 11 mM glucose, 2.5 mM KCl, 1.0 mM K_2HPO_4, 2.5 mM $CaCl_2$, 1.3 mM $MgCl_2$, 0.05 mM AP5, 0.02 mM DNQX and 0.008 mM DHβE. Recording electrodes of 3–5 MΩ were filled with an internal solution containing 110 mM $CsMeSO_4$, 2.8 mM NaCl, 4 mM EGTA, 5 mM TEA-Cl, 4 mM adenosine 5′-triphosphate (magnesium salt), 0.3 mM guanosine 5′-triphosphate (trisodium salt), 20 mM HEPES and 10 mM phosphocreatine (disodium salt), 0.025 mM Alexa 488 (for SACs) and 0.025 mM Alexa 594 (for nDSGCs), pH 7.2. Data were acquired using PCLAMP 10 recording software and a Multiclamp 700A amplifier (Molecular Devices). The GABAergic conductance was calculated from the linear portion of the current–voltage curve for the SAC-evoked currents in nDSGCs. After recording, we imaged the dye-filled SACs and the nDSGCs using a custom-modified two-photon microscope as described previously[30] (FluoView 300, Olympus America) at 745 nm. Images were acquired at z intervals of 0.5 μm using a ×60 objective (Olympus LUMPlanFl/IR ×60/0.90W). SAC and nDSGC processes were reconstructed from image stacks with NEUROLUCIDA. For *in vivo* injections, we anaesthetized animals with 3.5% isoflurane/2% O_2. The eyelid was then opened with fine forceps, and 1 μl of 10 mM muscimol, 500 μM gabazine or saline was injected using a fine glass micropipette. Injections were made with a picospritzer (World

Precision Instruments) generating 20-p.s.i., 3-ms-long positive pressure. We repeated this procedure every 48 h, starting at P6 and ending at P12. Whole-mount retina preparation and two-photon targeted recording for light responses was performed according to previously described techniques[30]. The direction selectivity index was computed as previously described[19].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 31 May; accepted 13 October 2010.
Published online 5 December 2010.

1. Euler, T., Detwiler, P. B. & Denk, W. Directionally selective calcium signals in dendrites of starburst amacrine cells. *Nature* **418,** 845–852 (2002).
2. Demb, J. B. Cellular mechanisms for direction selectivity in the retina. *Neuron* **55,** 179–186 (2007).
3. Fried, S. I., Munch, T. A. & Werblin, F. S. Mechanisms and circuitry underlying directional selectivity in the retina. *Nature* **420,** 411–414 (2002).
4. Barlow, H. B. & Levick, W. R. The mechanism of directionally selective units in rabbit's retina. *J. Physiol. (Lond.)* **178,** 477–504 (1965).
5. Oyster, C. W. The analysis of image motion by the rabbit retina. *J. Physiol. (Lond.)* **199,** 613–635 (1968).
6. Ariel, M. & Daw, N. W. Pharmacological analysis of directionally sensitive rabbit retinal ganglion cells. *J. Physiol. (Lond.)* **324,** 161–185 (1982).
7. Kittila, C. A. & Massey, S. C. Effect of ON pathway blockade on directional selectivity in the rabbit retina. *J. Neurophysiol.* **73,** 703–712 (1995).
8. Weng, S., Sun, W. & He, S. Identification of ON-OFF direction-selective ganglion cells in the mouse retina. *J. Physiol. (Lond.)* **562,** 915–923 (2005).
9. Famiglietti, E. V. Synaptic organization of starburst amacrine cells in rabbit retina: analysis of serial thin sections by electron microscopy and graphic reconstruction. *J. Comp. Neurol.* **309,** 40–70 (1991).
10. Stacy, R. C. & Wong, R. O. Developmental relationship between cholinergic amacrine cell processes and ganglion cell dendrites of the mouse retina. *J. Comp. Neurol.* **456,** 154–166 (2003).
11. Yoshida, K. *et al.* A key role of starburst amacrine cells in originating retinal directional selectivity and optokinetic eye movement. *Neuron* **30,** 771–780 (2001).
12. Amthor, F. R., Keyser, K. T. & Dmitrieva, N. A. Effects of the destruction of starburst-cholinergic amacrine cells by the toxin AF64A on rabbit retinal directional selectivity. *Vis. Neurosci.* **19,** 495–509 (2002).
13. Chen, Y. C. & Chiao, C. C. Symmetric synaptic patterns between starburst amacrine cells and direction selective ganglion cells in the rabbit retina. *J. Comp. Neurol.* **508,** 175–183 (2008).
14. Famiglietti, E. V. A structural basis for omnidirectional connections between starburst amacrine cells and directionally selective ganglion cells in rabbit retina, with associated bipolar cells. *Vis. Neurosci.* **19,** 145–162 (2002).
15. Jeon, C. J. *et al.* Pattern of synaptic excitation and inhibition upon direction-selective retinal ganglion cells. *J. Comp. Neurol.* **449,** 195–205 (2002).
16. Chan, Y. C. & Chiao, C. C. Effect of visual experience on the maturation of ON-OFF direction selective ganglion cells in the rabbit retina. *Vision Res.* **48,** 2466–2475 (2008).
17. Chen, M., Weng, S., Deng, Q., Xu, Z. & He, S. Physiological properties of direction-selective ganglion cells in early postnatal and adult mouse retina. *J. Physiol. (Lond.)* **587,** 819–828 (2009).
18. Elstrott, J. *et al.* Direction selectivity in the retina is established independent of visual experience and cholinergic retinal waves. *Neuron* **58,** 499–506 (2008).
19. Huberman, A. D. *et al.* Genetic identification of an On-Off direction-selective retinal ganglion cell subtype reveals a layer-specific subcortical map of posterior motion. *Neuron* **62,** 327–334 (2009).
20. Watanabe, D. *et al.* Ablation of cerebellar Golgi cells disrupts synaptic integration involving GABA inhibition and NMDA receptor activation in motor coordination. *Cell* **95,** 17–27 (1998).
21. Togashi, H. *et al.* Cadherin regulates dendritic spine morphogenesis. *Neuron* **35,** 77–89 (2002).
22. Zhu, H. & Luo, L. Diverse functions of N-cadherin in dendritic and axonal terminal arborization of olfactory projection neurons. *Neuron* **42,** 63–75 (2004).
23. Dong, W., Sun, W., Zhang, Y., Chen, X. & He, S. Dendritic relationship between starburst amacrine cells and direction-selective ganglion cells in the rabbit retina. *J. Physiol. (Lond.)* **556,** 11–17 (2004).
24. Vaney, D. I., Collin, S. P. & Young, H. M. in *Neurobiology Of The Inner Retina* (eds Weiler, R. & Osborne, N. N.) 157–168 (Springer, 1989).
25. Elstrott, J. & Feller, M. B. Direction-selective ganglion cells show symmetric participation in retinal waves during development. *J. Neurosci.* **30,** 11197–11201 (2010).
26. Wang, C. T. *et al.* GABA(A) receptor-mediated signaling alters the structure of spontaneous activity in the developing retina. *J. Neurosci.* **27,** 9130–9140 (2007).
27. Huang, Z. J. Activity-dependent development of inhibitory synapses and innervation pattern: role of GABA signalling and beyond. *J. Physiol. (Lond.)* **587,** 1881–1888 (2009).
28. Petreanu, L., Mao, T., Sternson, S. M. & Svoboda, K. The subcellular organization of neocortical excitatory connections. *Nature* **457,** 1142–1145 (2009).
29. Li, Y., Van Hooser, S. D., Mazurek, M., White, L. E. & Fitzpatrick, D. Experience with moving visual stimuli drives the early development of cortical direction selectivity. *Nature* **456,** 952–956 (2008).
30. Wei, W., Elstrott, J. & Feller, M. B. Two-photon targeted recording of GFP-expressing neurons for light responses and live-cell imaging in the mouse retina. *Nature Protocols* **5,** 1347–1352 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Nakanishi for *mGluR2–GFP* mice, A. Huberman for *Drd4–GFP* mice, J. Elstrott for help with MATLAB software, X. Han for mouse genotyping, J. Ledue for imaging assistance and A. Blankenship for reading the manuscript. This work was supported by grants RO1EY013528 and ARRA EY019498 from the National Institutes of Health.

Author Contributions W.W. conducted the electrophysiology and imaging experiments, and manuscript preparation; A.M.H. conducted intraocular injections, analysis of retinogeniculate projection patterns and manuscript preparation. K.Z. conducted NEUROLUCIDA reconstructions and analysis. M.B.F. was involved in the experimental design, data analysis of Supplementary Fig 3c–e and manuscript preparation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.B.F. (mfeller@berkeley.edu).

METHODS

Mice. *Drd4–GFP* mice in the Swiss Webster background were obtained from MMRRC[19] (http://www.mmrrc.org/strains/231/0231.html), and *mGluR2–GFP* mice were a gift from Shigatada Nakanishi, Osaka. Both strains were backcrossed to the C57BL/6 background in our laboratory. The *Drd4–GFP*/*mGluR2–GFP* double-transgenic mice were obtained by crossing the two single-transgenic lines.

Whole-cell patch-clamp recording. Single or dual whole-cell voltage-clamp recordings from SACs and nDSGCs were performed in oxygenated artificial cerebrospinal fluid at 32–34 °C containing 119.0 mM NaCl, 26.2 mM $NaHCO_3$, 11 mM glucose, 2.5 mM KCl, 1.0 mM K_2HPO_4, 2.5 mM $CaCl_2$, 1.3 mM $MgCl_2$, 0.05 mM AP5, 0.02 mM DNQX and 0.008 mM DHβE. Recording electrodes of 3–5 MΩ were filled with an internal solution containing 110 mM $CsMeSO_4$, 2.8 mM NaCl, 4 mM EGTA, 5 mM TEA-Cl, 4 mM adenosine 5′-triphosphate (magnesium salt), 0.3 mM guanosine 5′-triphosphate (trisodium salt), 20 mM HEPES and 10 mM phosphocreatine (disodium salt), 0.025 mM Alexa 488 (for SACs) and 0.025 mM Alexa 594 (for nDSGCs), pH 7.25. Data were acquired using PCLAMP 10 recording software and a Multiclamp 700A amplifier (Molecular Devices), filtered at 4 kHz and digitized at a sampling rate of 10 kHz. The GABAergic whole-cell conductance was calculated from the linear portion of the current–voltage curve for the SAC-evoked currents in nDSGCs and analysed using MATLAB software.

Two-photon targeted loose-patch recording of *GFP*-positive neurons for light response. *Drd4–GFP* mice were anaesthetized with isofluorane and decapitated in accordance with the UC Berkeley Institutional Animal Care and Use Committee and in conformance with the NIH Guide for the Care and Use of Laboratory Animals, the Public Health Service Policy and the SFN Policy on the Use of Animals in Neuroscience Research. Under infrared illumination, retinas were isolated from the pigment epithelium in oxygenated Ames' medium (Sigma), cut into dorsal and ventral halves, and mounted over a hole of 1–1.5 mm^2 on filter paper (Millipore) with the photoreceptor layer facing down. Retinas were kept in darkness at 25 °C in Ames' medium bubbled with 95% O2/5% CO2 until use (0–7 h). Recording electrodes of 3–5 MΩ were filled with Ames' medium. GFP fluorescence was detected with a custom-built, FluoView-based two-photon microscope and a Ti:sapphire laser (Coherent) tuned to 920 nm, a wavelength that minimally activates mouse photoreceptors and therefore preserves light response. GFP cells were then targeted for loose-patch recordings using PCLAMP 10 recording software and a Multiclamp 700A amplifier.

Visual stimuli were generated as previously described[19]. Briefly, a white, monochromatic organic light-emitting display (OLEDXL, eMagin; 800 × 600 pixel resolution, 85-Hz refresh rate) was controlled by an Intel Core Duo computer with a Windows XP operating system. Drifting square-wave gratings (spatial frequency, 225 μm per cycle; temporal frequency, 4 cycles s^{-1}; 30° s^{-1} in 12 pseudorandomly chosen directions spaced at 30 intervals, with each presentation lasting 3 s and followed by 500 ms of grey screen) were generated from the OLED using MATLAB and the Psychophysics Toolbox, and were projected through the ×60 water-immersion objective (LUMPlanFl/IR, NA 0.9) via the side port of the microscope, centred on the soma of the recorded cell and focused on the photoreceptor layer. Loose-patch recordings were obtained during the stimulus presentation and analysed using MATLAB. A detailed, step-by-step protocol of the two-photon targeted recording of light response can be found in ref. 30.

Two-photon microscopy and morphological reconstruction. After paired recording, the Alexa-488-filled SACs and the Alexa-594-filled nDSGCs in the *Drd4–GFP*/*mGluR2–GFP* mice were imaged using the two-photon microscope at 745 nm. At this wavelength, GFP is not efficiently excited but both Alexa 488 and Alexa 594 are brightly fluorescent. Therefore, the morphology of the Alexa-488-filled SACs could be distinguished from the very weak GFP fluorescence. Image stacks were acquired at z intervals of 0.5 μm and resampled three times for each stack using a ×60 objective (Olympus LUMPlanFl/IR ×60/0.90W), covering the entire dendritic fields of the SACs and nDSGCs. Image stacks from 25 SAC–nDSGC pairs were then imported into NEUROLUCIDA (MBF Biosciences) and reconstructed in three dimensions. The densities of contacts and cofasciculations were measured from the reconstructions.

Intraocular injections. *Drd4–GFP* animals were anaesthetized with 3.5% isoflurane/2% O_2. The eyelid was then opened with fine forceps, and 1 μl of 10 mM muscimol (Tocris), 500 μM gabazine (Tocris) or saline was injected using a fine glass micropipette. Injections were made with a picospritzer (World Precision Instruments) generating 20-p.s.i., 3-ms-long positive pressure. To prevent efflux of the injected solution, removal of the pipette tip from the eye was done slowly and gentle pressure was then applied to the injection site with a sterile cotton swab for ~10 s. This procedure was repeated every 48 h, starting at P6 and ending at P12.

Statistical analysis. Grouped data are presented as mean ± s.d. or s.e.m. as indicated. Data sets were tested for normality, and statistical differences were examined using one-way analysis of variance and *post hoc* comparisons using Student's *t*-test with Bonferroni corrections (MATLAB).

LETTER

doi:10.1038/nature09711

Spatially asymmetric reorganization of inhibition establishes a motion-sensitive circuit

Keisuke Yonehara[1], Kamill Balint[1], Masaharu Noda[2,3], Georg Nagel[4], Ernst Bamberg[5,6] & Botond Roska[1]

Spatial asymmetries in neural connectivity have an important role in creating basic building blocks of neuronal processing[1,2]. A key circuit module of directionally selective (DS) retinal ganglion cells is a spatially asymmetric inhibitory input from starburst amacrine cells[3–5]. It is not known how and when this circuit asymmetry is established during development. Here we photostimulate mouse starburst cells targeted with channelrhodopsin-2 (refs 6–8) while recording from a single genetically labelled type of DS cell[9,10]. We follow the spatial distribution of synaptic strengths between starburst and DS cells during early postnatal development before these neurons can respond to a physiological light stimulus, and confirm connectivity by monosynaptically restricted trans-synaptic rabies viral tracing. We show that asymmetry develops rapidly over a 2-day period through an intermediate state in which random or symmetric synaptic connections have been established. The development of asymmetry involves the spatially selective reorganization of inhibitory synaptic inputs. Intriguingly, the spatial distribution of excitatory synaptic inputs from starburst cells is significantly more symmetric than that of the inhibitory inputs at the end of this developmental period. Our work demonstrates a rapid developmental switch from a symmetric to asymmetric input distribution for inhibition in the neural circuit of a principal cell.

DS retinal ganglion cells respond to movement in a 'preferred' direction with robust spiking, but show minimal response to movement in the opposite, or 'null', direction[2,11–13]. DS cells receive GABAergic inhibitory inputs from starburst amacrine cell processes pointing in the null direction, but not from those pointing in the preferred direction[3,14]. Glutamatergic excitatory input from bipolar cells is also directionally selective. Interestingly, starburst cells also communicate to DS cells using acetylcholine[15,16], but this excitatory connection seems to be symmetric[14]. Directional selectivity is present before eye opening (around postnatal day 13 (P13) in mice), as well as in dark-reared animals[9,17–19], indicating that the establishment of circuit asymmetry does not require visual experience. How and when such highly specific synaptic connections are established between starburst and DS cells during development remain unknown. Retinal cells do not respond to light until P10–11 in mice[18,20], with the exception of melanopsin-containing ganglion cells[21], which limits the ability to follow the early development of functional connectivity. Directional selectivity may develop by the asymmetric refinement of previously formed inhibitory connections (Supplementary Fig. 1a) or, alternatively, the inhibitory synaptic inputs form asymmetrically (Supplementary Fig. 1b).

To distinguish between these possibilities, we probed the spatial distribution of synaptic strengths from starburst amacrine cells to individual ON DS cells during postnatal development. ON DS cells respond to slow movement and are critical for mediating the optokinetic reflex[22–24]. In SPIG1–GFP (SPIG1, also known as Fstl4, locus driving green fluorescent protein expression) knock-in mice, upward-motion-preferring ON DS cells are selectively labelled with GFP throughout development in most retinal regions[9,10].

To activate the starburst cells of SPIG1–GFP mice before amacrine and ganglion cells receive light-driven inputs from bipolar cells, this mouse line was crossed with another line expressing Cre recombinase specifically in starburst cells (choline acetyltransferase (ChAT)–Cre knock-in mice)[25]. At P0, we transduced these SPIG1–GFP × ChAT–Cre mice with a Cre-recombinase-dependent adeno-associated virus (AAV) carrying a reversed and double-floxed C128T mutant channelrhodopsin-2 (ChR2c)[6–8] followed by 2A–DsRed2 (ChR2c–2A–DsRed2, see Methods, Fig. 1a). 2A sequence codes for a *cis*-acting hydrolase element[26] that creates equimolar amounts of ChR2c and red-fluorescent, soluble DsRed2. A soluble marker in the cell body allowed easier quantification of fluorescence, and therefore ChR2 expression, than in a membrane-bound fusion construct. ChR2c-expressing cells are responsive to light at an intensity 50-fold lower than cells expressing wild-type ChR2 (ref. 7) and could, therefore, be activated by light patterns generated by an overhead projector.

Immunohistochemistry showed that all DsRed2-marked neurons were also positive for ChAT, a marker for starburst cells. Conversely, a substantial fraction (~60%) of starburst cells in both the ganglion cell layer (GCL) and inner nuclear layer (INL) (Fig. 1b, c) were DsRed2-labelled. Therefore, starburst cells, but no other cell type in these mice, are labelled red and express light-sensitive ChR2c, whereas upward-motion-preferring ON DS cells are labelled green (Fig. 1e). First, we characterized the light-excitability of ChR2c-positive starburst cells in intact, isolated retinas between P6 and P9. Light illumination evoked robust currents in DsRed2-expressing cells, even in the presence of glutamatergic synaptic blockers (CPP, NBQX, APB), suggesting ChR2c as the source of the currents (Fig. 1d). Increasing illumination evoked increasing membrane potential changes in starburst cells and, as expected due to the 2A element, the red fluorescence intensity of the recorded cell bodies correlated well ($R = 0.83$) with the magnitude of the membrane potential change at the stimulation intensity which is used to test the distribution of synaptic strengths in subsequent experiments (Supplementary Fig. 2).

To test whether the genetically tagged neural circuit could report the synaptic strengths from starburst to ON DS cells, we isolated excitatory and inhibitory inputs to ON DS cells at P8 while stimulating ChR2c-expressing starburst cells with light patterns (see Methods). A full-field light step elicited both inhibitory and excitatory currents in ON DS cells (Fig. 1f and Supplementary Fig. 3). The inhibitory component was blocked by the GABA receptor antagonist picrotoxin, and the excitatory input by the cholinergic receptor antagonist curare. Blocking glutamate receptors had no effect on the light-evoked currents (Fig. 1f) or on the miniature excitatory postsynaptic currents (mEPSC, Supplementary Fig. 4). mEPSCs were blocked by curare (Supplementary Fig. 4). These results confirmed that ON DS cells receive GABAergic and

[1]Neural Circuit Laboratories, Friedrich Miescher Institute for Biomedical Research, 4058 Basel, Switzerland. [2]Division of Molecular Neurobiology, National Institute for Basic Biology, 444-8787 Okazaki, Japan. [3]School of Life Science, The Graduate University for Advanced Studies, 444-8787 Okazaki, Japan. [4]Universität Würzburg, Botanik I, Julius-von-Sachs-Platz 2, 97082 Würzburg, Germany. [5]Max-Planck-Institut für Biophysik, Max-von-Laue Strasse 3, 60438 Frankfurt, Germany. [6]Johann Wolfgang Goethe-Universität, Institut für Biophysikalische Chemie, Max-von-Laue-Strasse 9, 60438 Frankfurt, Germany.

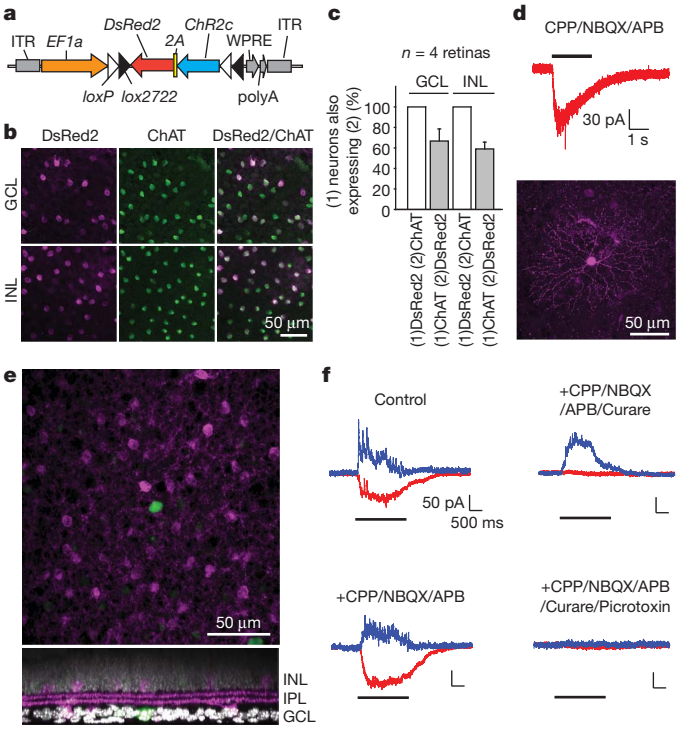


Figure 1 | Targeting of ChR2c to starburst amacrine cells at P8. **a**, AAV vector. EF1a, promoter; ITR, inverted terminal repeat; WPRE, woodchuck post-transcriptional regulatory element. **b**, Confocal images from an AAV-transduced retina. **c**, Relationship between DsRed2-expressing and ChAT-positive cells. **d**, Top, excitatory currents in an AAV-labelled starburst cell in the presence of synaptic blockers. Full-field flash stimulus. Bottom, confocal image of the recorded starburst cell. **e**, Top, confocal image of a retina in which ON DS cells are expressing GFP (green) and starburst amacrine cells are expressing ChR2c and DsRed2 (magenta). Bottom, side-view. IPL, inner plexiform layer. **f**, Synaptic currents recorded at −60 mV (red) and 20 mV (blue) holding potentials from a GFP-positive ON DS cell in response to a full-field flash. Error bars, s.d.

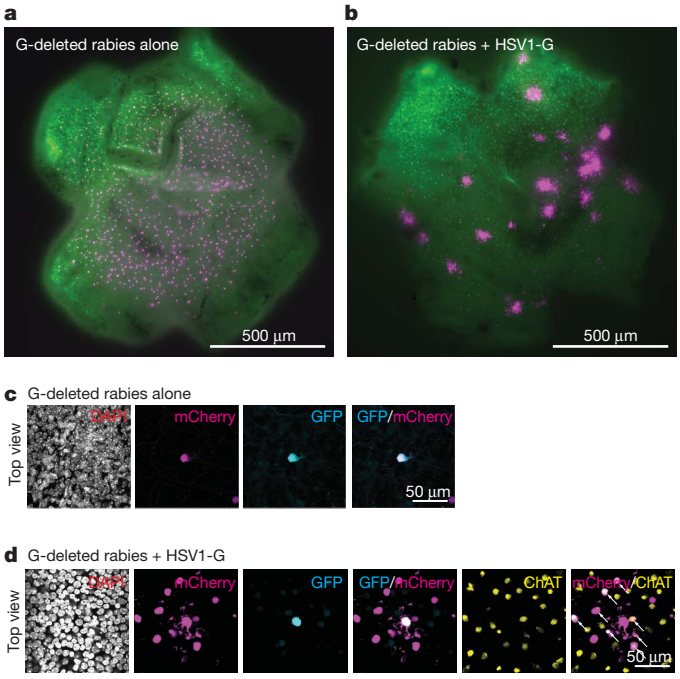


Figure 2 | Monosynaptically restricted circuit mapping initiated from ON DS cells. **a**, **b**, Live images of SPIG1–GFP retinas at P6 in which GFP-labelled ON DS cells (green) were infected with G-deleted rabies expressing mCherry (magenta) either alone (**a**) or in combination with G-encoding herpes virus (**b**). **c**, Confocal images of a GFP-labelled ON DS cell (cyan) infected with G-deleted rabies virus only (magenta). **d**, Confocal images of an ON DS cell infected with both G-deleted rabies virus and G-encoding herpes virus from **b**. Most of the labelled presynaptic cells (arrows) are ChAT-positive starburst cells (yellow) in the GCL.

cholinergic synaptic inputs in response to starburst cell stimulation, but do not receive glutamatergic synaptic input from bipolar cells at this stage.

The ChR2c-assisted synaptic strength mapping depends on direct connections between starburst and ON DS cells during early postnatal development. To test whether this is the case, we performed monosynaptically restricted retrograde synaptic tracing with G-deleted rabies virus[27,28] complemented with G-expressing herpes virus initiated from GFP-labelled ON DS cells (see Methods, Supplementary Figs 5 and 6). At P6, starburst cells were rabies-labelled around infected GFP-marked ON DS cells (Fig. 2), indicating that starburst cells are directly connected to ON DS cells at this developmental stage.

Having confirmed monosynaptic connection from starburst cells already at P6, we investigated the spatial distribution of the strength of synaptic connections by stimulating starburst cells with light steps in eight sectors surrounding the recorded ON DS cells (Fig. 3). We calculated a spatial asymmetry index (SAI) that quantified the degree of spatial asymmetry of the synaptic inputs to ON DS cells along the dorso-ventral axis (see Methods). We found that the inhibitory input was already spatially asymmetric along the dorso-ventral axis by P8; stimulation of the ventral (null) side evoked more inhibitory current than stimulation of the dorsal (preferred) side. In contrast, the excitatory input was significantly more symmetric along the same axis (Fig. 3 and Supplementary Fig. 7). To avoid potential bias due to non-uniform viral transduction we normalized the synaptic currents with either the number of DsRed2-expressing starburst cells (using a threshold) or the sum of the measured red fluorescence (which reflects the voltage change in starburst cells, as shown before, Supplementary Fig. 2) of the starburst cells in each of the eight sectors in which the light stimulus was presented (see Methods). The normalized responses, like the recorded raw responses, also showed asymmetric inhibition and more symmetric excitation along the dorso-ventral axis at P8 (Fig. 3, Supplementary Figs 7 and 8). In contrast to P8, the raw and normalized inhibition and excitation at P6 was symmetric along the same axis (Fig. 3, Supplementary Figs 7 and 8). The lack of asymmetry in inhibition at P6 was not due to ineffective activation of starburst cells because of low ChR2c expression level, since half-maximal activation at P9, which should be similar to maximal activation at P6 in terms of eliciting changes of membrane potential (Supplementary Fig. 2), revealed asymmetry (Supplementary Fig. 9) .

Next, we investigated the emergence of asymmetry from P6 to P9 (Fig. 4). SAI of inhibition increased significantly, but there was no significant change in excitation between any pairs of days (Supplementary Fig. 8, note: the lack of stars between pairs of conditions on any of the figures means no significant change). The lack of statistically significant change in excitation was not due to saturating intensities because half-maximal activation of excitatory inputs to ON DS cells did not significantly change the SAI of excitation at P9 (Supplementary Fig. 9). The mean direction of inhibitory input of individual recorded cells, computed as the vector sum of inputs for all eight directions, was random at P6 but became confined to the ventral side by P8 (direction of red bars in Fig. 4h). Because the variation in DsRed2 expression across the eight sectors was not statistically different between P6 and P9 (Supplementary Fig. 10), the randomness at P6 is not due to greater variation in gene expression from AAV at earlier time points. We conclude that before P6 the spatial distribution of inhibitory connectivity between starburst cells and ON DS cells is either random or symmetric (with some synapses having little strength). Inhibitory connectivity rapidly reorganizes to become asymmetric along the 'preferred-null' (dorso-ventral) axis between P6 and P8, whereas excitatory cholinergic

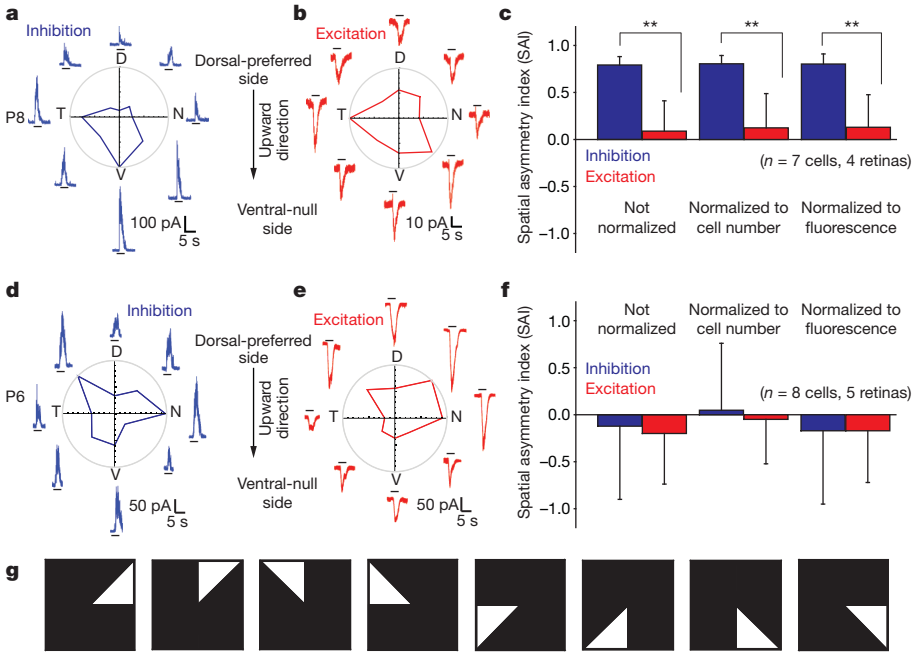


Figure 3 | ChR2c-assisted circuit mapping at P8 and P6. **a**–**f**, Recordings from ON DS cells at P8 (**a**–**c**) and P6 (**d**–**f**). Inhibitory (**a**, **d**) and excitatory (**b**, **e**) postsynaptic currents elicited in an ON DS cell by the stimulation of eight sectors surrounding the cell. Polar plots are also shown. **c**, **f**, Spatial asymmetry index (SAI) for inhibition and excitation. **g**, Sketch of light patterns used to stimulate one of eight sectors around the recorded ON DS cell. Error bars, s.d.

input remains significantly more symmetric throughout this developmental period (Supplementary Fig. 1a).

Is inhibitory connectivity strengthened at the ventral ('null') and weakened at the dorsal ('preferred') side or is only one of these two mechanisms driving the development of asymmetry? Since starburst cells similarly control the strength of cholinergic excitation and GABAergic inhibition to ON DS cells (Supplementary Fig. 11) and excitation is not significantly different along the dorso-ventral axis from P6 to P9, the ratio of inhibition to excitation (neither normalized) in the dorsal and ventral sides should be a measure of the inhibitory synaptic strength that depends less on the level of ChR2c expression (which increases over the days, Supplementary Fig. 2d) than inhibition alone. This ratio increased in the ventral (though the increase was not significant) and decreased in the dorsal side (Supplementary Fig. 12), suggesting that 'push-pull' synapse reorganization is at work.

The retinal stratum in which ON DS cells extend their dendrites embodies three different directionally selective computations that lead to preferential responses to nasal, upward and downward motion in different types of ON DS cells. We suggest that, in the physical space shared by these circuits, it is the spatially selective refinement of the distribution of inhibitory input strength to each DS ganglion cell that underlies the establishment of each directionally selective retinal circuit.

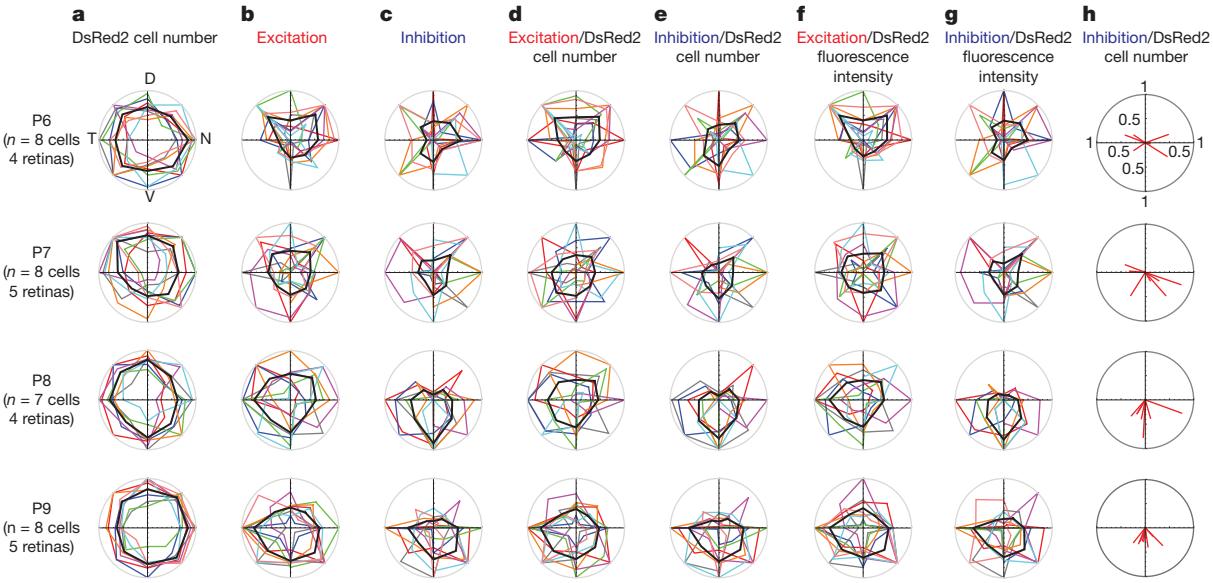


Figure 4 | Development of asymmetry. Coloured lines indicate normalized current responses from individual cells and the black line indicates the mean response of all recorded cells (or the mean cell number for **a**). Polar plot of the number of DsRed2-expressing cells (**a**), excitatory (**b**) and inhibitory (**c**) inputs, excitatory (**d**) and inhibitory (**e**) inputs normalized to the number of DsRed2 expressing cells in each sector, excitatory (**f**) and inhibitory (**g**) inputs normalized to the mean DsRed2 fluorescence intensity of cells in each sector. **h**, Red bars indicate the vector sum of inhibitory inputs normalized to the number of DsRed2 expressing cells in each sector.

Higher-order brain computations, for example orientation selectivity in the visual cortex, also rely on spatial circuit asymmetries. Mechanistic insights from the development of retinal directional selectivity may help to understand how asymmetry in cortical circuits is established.

METHODS SUMMARY

On the day of birth (P0) starburst amacrine cells in the progeny of a cross between mice expressing Cre in starburst amacrine cells and SPIG1–GFP mice expressing GFP in ON DS cells were labelled with ChR2c *in vivo* by transduction with a Cre-recombinase-dependent AAV. Retinas were isolated at P6 or later and GFP-labelled ON DS cells were recorded in voltage clamp at −60 mV (for excitation) or 20 mV (for inhibition)[29] guided by a two-photon microscope at 930 nm (ref. 30). Photostimulation was performed with white light steps (duration 2 s, inter-stimulus interval 10 s) generated by a digital light projector (PLUS Vision) and directed onto each of eight sectors surrounding the recorded cell. Stereotaxic injections of rabies and herpes viruses to the medial terminal nucleus (MTN) of SPIG1–GFP mice were performed at P1 and infected retinas were isolated at P6.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 July; accepted 2 December 2010.
Published online 19 December 2010.

1. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.)* **160,** 106–154 (1962).
2. Barlow, H. B. & Hill, R. M. Selective sensitivity to direction of movement in ganglion cells of the rabbit retina. *Science* **139,** 412–414 (1963).
3. Fried, S. I., Munch, T. A. & Werblin, F. S. Mechanisms and circuitry underlying directional selectivity in the retina. *Nature* **420,** 411–414 (2002).
4. Euler, T., Detwiler, P. B. & Denk, W. Directionally selective calcium signals in dendrites of starburst amacrine cells. *Nature* **418,** 845–852 (2002).
5. Lee, S. & Zhou, Z. J. The synaptic mechanism of direction selectivity in distal processes of starburst amacrine cells. *Neuron* **51,** 787–799 (2006).
6. Bamann, C., Gueta, R., Kleinlogel, S., Nagel, G. & Bamberg, E. Structural guidance of the photocycle of channelrhodopsin-2 by an interhelical hydrogen bond. *Biochemistry* **49,** 267–278 (2010).
7. Berndt, A., Yizhar, O., Gunaydin, L. A., Hegemann, P. & Deisseroth, K. Bi-stable neural state switches. *Nature Neurosci.* **12,** 229–234 (2008).
8. Radu, I. *et al.* Conformational changes of channelrhodopsin-2. *J. Am. Chem. Soc.* **131,** 7313–7319 (2009).
9. Yonehara, K. *et al.* Identification of retinal ganglion cells and their projections involved in central transmission of information about upward and downward image motion. *PLoS ONE* **4,** e4320 (2009).
10. Yonehara, K. *et al.* Expression of SPIG1 reveals development of a retinal ganglion cell subtype projecting to the medial terminal nucleus in the mouse. *PLoS ONE* **3,** e1533 (2008).
11. Vaney, D. I. & Taylor, W. R. Direction selectivity in the retina. *Curr. Opin. Neurobiol.* **12,** 405–410 (2002).
12. Barlow, H. B. & Levick, W. R. The mechanism of directionally selective units in rabbit's retina. *J. Physiol. (Lond.)* **178,** 477–504 (1965).
13. Demb, J. B. Cellular mechanisms for direction selectivity in the retina. *Neuron* **55,** 179–186 (2007).
14. Fried, S. I., Munch, T. A. & Werblin, F. S. Directional selectivity is formed at multiple levels by laterally offset inhibition in the rabbit retina. *Neuron* **46,** 117–127 (2005).
15. Ariel, M. & Daw, N. W. Pharmacological analysis of directionally sensitive rabbit retinal ganglion cells. *J. Physiol. (Lond.)* **324,** 161–185 (1982).
16. Masland, R. H. & Ames, A. III. Responses to acetylcholine of ganglion cells in an isolated mammalian retina. *J. Neurophysiol.* **39,** 1220–1235 (1976).
17. Chan, Y. C. & Chiao, C. C. Effect of visual experience on the maturation of ON-OFF direction selective ganglion cells in the rabbit retina. *Vision Res.* **48,** 2466–2475 (2008).
18. Chen, M., Weng, S., Deng, Q., Xu, Z. & He, S. Physiological properties of direction-selective ganglion cells in early postnatal and adult mouse retina. *J. Physiol. (Lond.)* **587,** 819–828 (2009).
19. Elstrott, J. *et al.* Direction selectivity in the retina is established independent of visual experience and cholinergic retinal waves. *Neuron* **58,** 499–506 (2008).
20. Tian, N. & Copenhagen, D. R. Visual deprivation alters development of synaptic function in inner retina after eye opening. *Neuron* **32,** 439–449 (2001).
21. Hattar, S., Liao, H. W., Takao, M., Berson, D. M. & Yau, K. W. Melanopsin-containing retinal ganglion cells: architecture, projections, and intrinsic photosensitivity. *Science* **295,** 1065–1070 (2002).
22. Oyster, C. W., Takahashi, E. & Collewijn, H. Direction-selective retinal ganglion cells and control of optokinetic nystagmus in the rabbit. *Vision Res.* **12,** 183–193 (1972).
23. Simpson, J. I. The accessory optic system. *Annu. Rev. Neurosci.* **7,** 13–41 (1984).
24. Sun, W., Deng, Q., Levick, W. R. & He, S. ON direction-selective ganglion cells in the mouse retina. *J. Physiol. (Lond.)* **576,** 197–202 (2006).
25. Ivanova, E., Hwang, G. S. & Pan, Z. H. Characterization of transgenic mouse lines expressing Cre recombinase in the retina. *Neuroscience* **165,** 233–243 (2010).
26. Tang, W. *et al.* Faithful expression of multiple proteins via 2A-peptide self-processing: a versatile and reliable method for manipulating brain circuits. *J. Neurosci.* **29,** 8621–8629 (2009).
27. Wickersham, I. R. *et al.* Monosynaptic restriction of transsynaptic tracing from single, genetically targeted neurons. *Neuron* **53,** 639–647 (2007).
28. Stepien, A. E., Tripodi, M. & Arber, S. Monosynaptic rabies virus reveals premotor network organization and synaptic specificity of cholinergic partition cells. *Neuron* **68,** 456–472 (2010).
29. Roska, B. & Werblin, F. Vertical interactions across ten parallel, stacked representations in the mammalian retina. *Nature* **410,** 583–587 (2001).
30. Münch, T. A. *et al.* Approach sensitivity in the retina processed by a multifunctional neural circuit. *Nature Neurosci.* **12,** 1308–1316 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank B. G. Scherf, S. Djaffer and J. Jüttner for technical assistance, T. Szikra for helping with light intensity calibration, V. Busskamp for suggesting the use of the 2A element and the cloning strategy for the 2A-based expression system, and V. Busskamp, K. Farrow, S. Oakeley and P. King for their comments on the manuscript. We thank E. Callaway for providing the rabies viruses and K. Conzelmann and S. Arber for discussion about rabies viruses. The study was supported by the Friedrich Miescher Institute for Biomedical Research, a US Office of Naval Research Naval International Cooperative Opportunities in Science and Technology Program grant, a Marie Curie Excellence grant, a National Centre of Competence in Research Frontiers in Genetics grant, an European Research Council as well as RETICIRC, TREATRUSH and OPTONEURO grants from the European Union to B.R. and an EMBO Long-Term Fellowship to K.Y.

Author Contributions K.Y. performed and designed all retinal experiments, *in vivo* injection experiments with rabies, herpes and AAV viruses, developed all plasmids, analysed data and wrote the paper. K. B. grew and titred rabies viruses. M.N. developed SPIG1–GFP mice. G.N. and E.B. developed ChR2c. B.R. designed experiments, analysed data and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to B.R. (botond.roska@fmi.ch).

METHODS

Animals. ChAT–Cre mice were purchased from Jackson Laboratory (strain: B6;129S6-*Chat*$^{tm1(cre)Lowl}$/J). In ChAT–Cre mice, Cre recombinase is expressed under the control of the ChAT locus. In SPIG1–GFP mice GFP is expressed under the control of the SPIG1 locus[9,10]. To obtain neonatal SPIG1–GFP × ChAT–Cre mice we crossed SPIG1–GFP homozygous mice with ChAT–Cre homozygous mice. All animal procedures were performed in accordance with standard ethical guidelines (European Communities Guidelines on the Care and Use of Laboratory Animals, 86/609/EEC) and were approved by the Veterinary Department of the Canton of Basel-Stadt, Switzerland.

AAV plasmids. In the present paper, we refer to the C128T mutant ChR2 (refs 6–8) as ChR2c. To obtain pAAV-EF1a-double floxed-ChR2c–2A–DsRed2-WPRE-hGHpA we linearized pAAV-EF1a-double floxed-hChR2(H134R)-EYFP-WPRE-hGHpA (provided by K. Deisseroth) using NheI/AscI. ChR2c was PCR-amplified from pGEMHE-ChR2c using a HindIII-2A-covering primer. Forward primer: 5′-GCTAGCGCTAGCCACCATGGATTATGGAGGCGCCC TG-3′. Reverse primer: 5′-TCTCCCGCAAGCTTAAGAAGGTCAAAATTCTT GCCGGTGCCCTTGTTGAC-3′. DsRed2 was PCR-amplified from pDsRed2-N1 (Clontech) using a HindIII-2A-covering primer. Forward primer: 5′-ACCTTC TTAAGCTTGCGGGAGACGTCGAGTCCAACCCTGGGCCCATGGCCTCCT CCGAGAACGTC-3′. Reverse primer: 5′-GGCGCGCCGGCGCGCCCTATC ACAGGAACAGGTGGTGGCG-3′. These two PCR products were then digested with NheI, AscI and HindIII and triple ligation was performed.

AAV production. Serotype 7 recombinant AAVs were made by Penn Vector Core. Penn Vector Core performed the genome copy (GC) number titration (titre: 5.78×10^{12} GC per ml) using real-time PCR (TaqMan reagents, Applied Biosystems).

Logic of AAV labelling. We used the viral vector AAV EF1a-double floxed-ChR2c–2A–DsRed2 to target the expression of ChR2c to starburst cells expressing Cre. In AAV EF1a-double floxed-ChR2c–2A–DsRed2, two incompatible *loxP* variants[31], *loxP* and *lox2722*, flank an inverted version of *ChR2c* followed by the red fluorescent marker *DsRed2*. In the presence of Cre, a stochastic recombination of either *loxP* variant takes place, resulting in the inversion of *ChR2c–2A–DsRed2* into the sense direction, followed by expression of the ChR2c–2A–DsRed2. The *2A* element[32] codes for a *cis*-acting hydrolase element[26] that creates equimolar amounts of ChR2c and red-fluorescent DsRed2.

AAV injection. We injected the virus at the day of birth. SPIG1–GFP × ChAT–Cre mice were anesthetized with crushed ice. Virus (1.5 µl, 8.68×10^9 GC) was loaded into pulled glass pipettes (tip diameter, 30 µm) and injected into the vitreal space of both eyes using a microinjector (Narishige). After 6 days, DsRed2 expression was brightly detectable; hence all recordings were performed on retinas at P6 or later.

Preparation of retinas. Neonatal mice were killed by decapitation. Eyes were enucleated. The retinas were isolated and the pigment epithelium removed in Ringer's medium (in mM: 110 NaCl, 2.5 KCl, 1 $CaCl_2$, 1.6 $MgCl_2$, 10 D-glucose, 22 $NaHCO_3$, bubbled with 5% CO_2/95% O_2, pH 7.4) and mounted ganglion-cell-side up on a filter (MF-membrane, Millipore) with a 2-mm rectangular aperture in the centre. Before starting superfusion, DsRed2-positive regions together with GFP-positive cells were located with an epifluorescence stereomicroscope (Olympus) and photographed for later determination of the average DsRed2 fluorescence intensity around recorded ON DS cells for data normalization and orientation of the retina. Only GFP cells surrounded by DsRed2 expression were chosen for the recordings. The orientation of the isolated SPIG1–GFP × ChAT–Cre retina was determined by the pattern of GFP expression. In most retinal regions, GFP is expressed exclusively in one type of ON DS cells during the developmental period. An exception is the dorsal (slightly temporal) region, where GFP is expressed in many different ganglion and amacrine cell types. A thick axon bundle in this region runs in the dorso-ventral direction towards the optic disk and can be used as a compass in isolated retinas[9]. The retinas were superfused in Ringer's medium at 35–36 °C in the microscope chamber for the duration of the experiment. In this retinal preparation, ChR2c-mediated light responses could be measured for more than 8 h.

Two-photon imaging, electrophysiology and pharmacology. Fluorescent cells were found with the help of a two-photon microscope equipped with a Mai Tai HP two-photon laser (930 or 1,010 nm) (Spectra Physics) integrated into the electrophysiological setup[30]. Current recordings were made in whole-cell voltage clamp mode using an Axon Multiclamp 700B amplifier with borosilicate glass electrodes (BF100-50-10, Sutter Instruments) pulled to 7–9 MΩ, and filled with (in mM) 112.5 $CsCH_3SO_3$, 1 $MgSO_4$, 7.8×10^{-3} $CaCl_2$, 0.5 BAPTA, 10 HEPES, 4 ATP-Na_2, 0.5 GTP-Na_3, 5 lidocaine N-ethylbromide (Qx314-Br), 7.75 neurobiotin chloride, pH 7.2. Excitatory and inhibitory synaptic currents ('excitation' and 'inhibition', respectively) were separated by voltage-clamping the cell to the equilibrium potential of chloride (−60 mV) and unselective cation channels (20 mV), respectively[29]. For recording mEPSCs, cells were voltage-clamped at −60 mV and recorded for 3–5 min. Voltage recordings from DsRed2-positive starburst cells were made in whole-cell current clamp mode with glass electrodes pulled to 7–9 MΩ and filled with (in mM) 115 K-gluconate, 9.7 KCl, 1 $MgCl_2$, 0.5 $CaCl_2$, 1.5 EGTA, 10 HEPES, 4 ATP-Na_2,, 0.5 GTP-Na_3, pH 7.2. In pharmacological experiments, agents were bath-applied at the following concentrations: 10 µM CPP ((±)-3-(2-carboxypiperazin-4-yl) propyl-1-phosphonic acid, blocking NMDA receptors), 10 µM NBQX (6-nitro-2,3-dioxo-1,4-dihydrobenzo[f]quinoxaline-7-sulfonamide, blocking AMPA and kainate receptors), 10 µM APB (L-(+)-2-amino-4-phosphonobutyric acid, blocking metabotropic glutamate receptors and therefore blocking the ON pathway), 50 µM curare (tubocurarine chloride, blocking nicotinic acetylcholine receptors), 100 µM picrotoxin (blocking GABA A and C receptors). All chemicals were obtained from Sigma, with the exception of APB (Calbiochem). Data were analysed offline with Mathematica (Wolfram Research).

Photostimulation. ChR2c was activated with light generated by a digital light projector (V-332, PLUS Vision). The stimulation was generated via custom-made software (Matlab, Mathworks; Labview, National Instruments). Light intensity was measured using a spectrometer (USB 4000, Ocean Optics) calibrated with a reference source (LS1-Cal, Ocean Optics). Light intensity was modulated by using different grey scales (0–255) combined with different neutral density filters (ND0, ND10, ND20 and ND 30). To correlate the stimulus intensity and the change in membrane potential in ChR2c-expressing starburst cells (Supplementary Fig. 2), we used full-field flash at 24 different intensities (12.08–15.86 in Log intensity photons $cm^{-2}\,s^{-1}$) presented for 2 s with an inter-stimulus interval of 5 s. For the stimulation of the eight sectors (Figs 3, 4), each stimulus was presented for 2 s with an inter-stimulus interval of 10 s. The eight sectors were stimulated in a random order. The light pattern was focused on the GCL. To find a 'weak light intensity' that evoked half the maximum excitatory current input to ON DS cell (Supplementary Fig. 9), full-field flash at 24 different intensities (12.08–15.86 in Log intensity photons $cm^{-2}\,s^{-1}$) was presented sequentially (presentation for 2 s with an inter-stimulus interval of 5 s) initially and next the retinas were then stimulated in eight sectors using the determined light intensity.

Data collection and analysis. Light stimulation of each of the eight sectors (Fig. 3g) around the recorded ON DS cell body was repeated in each recorded ON DS cell 3–10 times for both excitation and inhibition and the mean light responses were determined for all eight directions. To correct for non-uniform viral expressions in the eight sectors, we performed two different types of normalizations. In the first procedure, we normalized the current evoked in each sector by the number of DsRed2-expressing cells in the sector within 200 µm of recorded ON DS cell bodies in the GCL. Here we used an arbitrary fluorescence threshold that was constant between experiments. We choose the particular distance of 200 µm because the radius of the dendrites of ON DS cells plus the radius of the processes of starburst amacrine cells at P6 and P9 were together less than 200 µm (data not shown).

In the second procedure, we normalized each sector to the average fluorescence intensity of the starburst cells in the sector and not just the number. This was reasonable because the fluorescence intensity of the starburst cells correlated well ($R = 0.83$) with the magnitude of the voltage response of the starburst cell at the stimulation intensity used for the mapping procedure (Supplementary Fig. 2): therefore, the average fluorescence is a measure of the stimulation strength of the starburst cells in the sector.

To yield the eight quantities plotted on polar plots, these normalized (or not normalized) values were further normalized to the largest of the eight numbers (for excitation and inhibition, independently). Note that this normalization is useful for eliminating variations in synaptic currents arising from the patch-clamp technique including series resistance and leak conductance. Note that the largest value (of the eight) to which normalization is performed is a mean of a distribution since each segment was stimulated 3–10 times (see above). The direction of the tuning was determined by multiplying the eight values above with the corresponding unit vectors pointing in eight directions and then forming the vector sum. The direction of the vector sum was interpreted as the direction of tuning.

The spatial asymmetry index (SAI) was calculated as:

$$\mathrm{SAI} = (I_{\mathrm{ventral}} - I_{\mathrm{dorsal}})/(I_{\mathrm{ventral}} + I_{\mathrm{dorsal}})$$

where I_{ventral} and I_{dorsal} are the amplitudes of the normalized or not normalized currents evoked by the stimulation of ventral or dorsal sectors (both normalized and not normalized SAIs are shown in Fig. 3 and Supplementary Figs 7–9).

Monosynaptic restriction of circuit tracing. To create rabies G-expressing replication-defective herpes simplex virus-1 (HSV1), the GFP open reading frame (ORF) in the HSV1 vector pR19EF1a-GFP-WCm (Biovex) was replaced with that of G. First, the G ORF was amplified by PCR from pHCMV-G[33] using primers 5′-GTGTCGTGAGGAATTCGTACCGGATCCTCTAGGCCACC-3′ and 5′-CC GCTTTACTTGTACATTACAGTCTGGTCTCACCCCCACT-3′. The GFP ORF

was removed from pR19EF1a-GFP-WCm by EcoRI/BsrGI digestion and the PCR fragment of G was recombined into EcoRI–BsrGI site using an in-fusion PCR kit (Takara-Clontech). The viral particles were produced by Biovex. G-deleted rabies virus encoding mCherry (SADΔG-mCherry)[34] was a gift from E. Callaway. Rabies virus expressing mCherry instead of the G glycoprotein was harvested from BHK-B19G cells (provided by E. Callaway) and centrifuged as described earlier[27]. We performed stereotaxic surgery to label ganglion cells projecting to the medial terminal nucleus (MTN). A cocktail of 10^3 plaque-forming units of rabies virus and 6×10^4 plaque-forming units of HSV1 in 20 nl DMEM were loaded into pulled-glass pipettes (tip inner diameter of 20–30 μm) and injected into the MTN using a microinjector (Narishige, IM-9B). For control experiments, we injected 2×10^5 plaque-forming units of rabies virus. Injections were performed with 24 mice at P1. Retinas were isolated at P6. Six well infected retinas at P6 were fixed by PFA and stained with antibodies. Brains were also isolated and the injection sites were localized. All rabies and HSV1 work was carried out under Biosafety level 2 conditions.

The key point of viral tracing was to infect with rabies and herpes viruses an upward-motion sensitive ON DS ganglion cell to initiate the retrograde passage of rabies viruses from this cell type. Because GFP exclusively labelled upward-motion sensitive ON DS cells in the ventral retina of SPIG1–GFP mice it was enough to examine the rabies-labelled cells around a GFP-labelled ganglion cell regardless of the injection site. The reason we performed the viral tracing in the SPIG1–GFP line was to have an internal control for the ganglion cell type for which we examine its local circuit. The fact that rabies did label GFP cells (in red) shows that the injection reached the MTN (because SPIG1–GFP cells exclusively target MTN[9,10], see also Supplementary Fig. 6). Even if by mistake we had injected these viruses also to other retinorecipient brain regions, this would not compromise our tracing results of the GFP-labelled ganglion cells provided, first, that the rabies-labelled circuits in the retina were far away so that the ganglion cell to which an amacrine cell is presynaptic to could be determined and, second, that only one ganglion cell was labelled in that local circuit. The reason we used low rabies titres for the herpes/rabies co-injections was to make sure that the circuits analysed were far away from each other in the retina (see Fig. 2b). In all circuits we analysed there was only one ganglion cell in it (which was GFP-labelled, see Fig. 2d). The definition of 'ganglion cell' was based on the existence of an axon.

Immunohistochemistry. After the experiments, retinas were fixed for 30 min in 4% (w/v) paraformaldehyde in PBS (137 mM NaCl, 2.7 mM KCl, 4.3 mM Na_2HPO_4, 1.47 mM KH_2PO_4, pH 7.4) and washed with PBS for at least 1 day at 4 °C. To aid penetration of the antibodies, retinas were frozen and thawed three times after cryoprotection with 30% (w/v) sucrose. All other procedures were carried out at room temperature. After washing in PBS, retinas were blocked for 1 h in 10% (w/v) normal donkey serum (NDS; Chemicon), 1% (w/v) bovine serum albumin (BSA), and 0.5% (v/v) Triton X-100 in PBS. Primary antibodies were incubated for 7 days in 3% (v/v) NDS, 1% (w/v) BSA, 0.02% (w/v) sodium azide and 0.5% (v/v) Triton X-100 in PBS. Secondary antibodies were incubated for 2 h in 3% (v/v) NDS, 1% (w/v) BSA, and 0.5% (v/v) Triton X-100 in PBS together with streptavidin-Alexa Fluor 633 (Invitrogen, 1:200) and DAPI (4′,6-diamidine-2-phenylindole dihydrochloride, Roche Diagnostics, 10 μg ml^{-1}) in some experiments. Streptavidin binds to neurobiotin and thus labels neurobiotin-filled cells; DAPI binds to DNA and therefore labels nuclei. After a final wash in PBS, retinas were embedded in Prolong Gold antifade (Invitrogen).

The following set of primary and secondary antibodies combinations were used in experiments in which we recorded from SPIG1 cells while stimulating ChR2c–2A–DsRed2-expressing starburst cells: (1) Primary: goat anti-ChAT (1:200, AB144P, Chemicon). Secondary: donkey anti-goat IgG conjugated with Alexa Fluor 405 (1:200, Invitrogen). (2) Primary: rabbit anti-red fluorescent protein (RFP; 1:200, AB3216, Chemicon). This primary antibody binds to DsRed2. Secondary: donkey anti-rabbit IgG conjugated with Cy3 (1:200, Jackson). (3) Primary: rat anti-GFP (1:500, 04404-84, Nacalai). Secondary: donkey anti-rat IgG conjugated with Alexa Fluor 488 (1:200, Invitrogen). The following set of primary and secondary antibodies combinations were used for staining rabies virus-infected retinas: (1) Primary: goat anti-ChAT (1:200, AB144P, Chemicon). Secondary: donkey anti-goat IgG conjugated with Alexa Fluor 633 (1:200, Invitrogen). (2) Primary: rabbit anti-red fluorescent protein (RFP; 1:200, AB3216, Chemicon). Secondary: donkey anti-rabbit IgG conjugated with Cy3 (1:200, Jackson). (3) Primary: rat anti-GFP (1:500, 04404-84, Nacalai). Secondary: donkey anti-rat IgG conjugated with Alexa Fluor 488 (1:200, Invitrogen).

Confocal analysis. Stained retinas were analysed with a Zeiss LSM 700 confocal microscope. The DsRed2-expressing cell number was assessed from *z*-stack images by using a ×20 lens, numerical aperture (NA) 0.7, ×0.5 digital zoom. All images were recorded at the same laser power and gain control. Overall morphologies of the recorded starburst or ganglion cells were assessed using a ×40 oil immersion lens, NA 1.2, ×0.5 digital zoom or ×63 oil immersion lens, NA 1.3, ×0.5 digital zoom. The mCherry-labelled presynaptic circuits of ON DS cells were assessed from *z*-stack images using a ×63 oil immersion lens, NA 1.3.

Statistical analysis. The non-parametric Mann–Whitney U test was used to compare data obtained from different cells on different days and the Wilcoxon signed rank test for comparing pairs of data where each pair was obtained from the same cell (excitation and inhibition). Significance is denoted by * for $P < 0.05$ and ** for $P < 0.01$. The error bars and ± values represent standard deviations (s.d.). On each figure, the lack of stars between any pairs of data signifies $P > 0.05$ and, therefore, that the two distributions are not statistically different.

31. Atasoy, D., Aponte, Y., Su, H. H. & Sternson, S. M. A. FLEX switch targets channelrhodopsin-2 to multiple cell types for imaging and long-range circuit mapping. *J. Neurosci.* **28,** 7025–7030 (2008).
32. Trichas, G., Begbie, J. & Srinivas, S. Use of the viral 2A peptide for bicistronic expression in transgenic mice. *BMC Biol.* **6,** 40 (2008).
33. Sena-Esteves, M., Tebbets, J. C., Steffens, S., Crombleholme, T. & Flake, A. W. Optimized large-scale production of high titer lentivirus vector pseudotypes. *J. Virol. Methods* **122,** 131–139 (2004).
34. Marshel, J. H., Mori, T., Nielsen, K. J. & Callaway, E. M. Targeting single neuronal networks for gene expression and cell labeling in vivo. *Neuron* **67,** 562–574 (2010).

LETTER

doi:10.1038/nature09641

Identification of two genes causing reinforcement in the Texas wildflower *Phlox drummondii*

Robin Hopkins[1] & Mark D. Rausher[1]

Species formation generates biological diversity and occurs when traits evolve that prevent gene flow between populations. Discerning the number and distribution of genes underlying these traits and, in a few cases, identifying the genes involved, has greatly enhanced our understanding over the past 15 years of species formation (reviewed by Noor and Feder[1] and Wolf *et al.*[2]). However, this work has almost exclusively focused on traits that restrict gene flow between populations that have evolved as a by-product of genetic divergence between geographically isolated populations. By contrast, little is known about the characteristics of genes associated with reinforcement, the process by which natural selection directly favours restricted gene flow during the formation of species. Here we identify changes in two genes that appear to cause a flower colour change in *Phlox drummondii*, which previous work has shown contributes to reinforcement. Both changes involve *cis*-regulatory mutations to genes in the anthocyanin biosynthetic pathway (ABP). Because one change is recessive whereas the other is dominant, hybrid offspring produce an intermediate flower colour that is visited less by pollinators, and is presumably maladaptive. Thus genetic change selected to increase prezygotic isolation also appears to result in increased postzygotic isolation.

Natural selection can directly favour species formation through a process termed reinforcement. If two incipient species experience secondary contact and produce maladaptive hybrids, selection favours decreased gene flow and increased reproductive isolation between them[3–5]. Reinforcement can be recognized by a resulting pattern known as reproductive character displacement: reproductive isolation is greater in sympatry than in allopatry[4]. Although the occurrence of reinforcement was historically controversial[6,7], empirical studies have documented reinforcing selection and reproductive character displacement in birds, insects, amphibians, plants and mammals, suggesting that it is a common step in the process of species formation (reviewed by Ortiz-Barrientos *et al.*[8] and Pfennig and Pfennig[9]). Despite this work, little is known about genetic changes associated with reinforcement (but see studies of quantitative trait loci by Ortiz-Barrientos *et al.*[10] and Saether *et al.*[11]).

Divergence of floral colour in *P. drummondii* constitutes one of the best-documented cases of reinforcement in plants[12] and exhibits the classic pattern of reproductive character displacement. *P. drummondii* and the closely related *P. cuspidata* produce similar light-blue flowers throughout the allopatric parts of their ranges. However, in the area of sympatry, *P. drummondii* has dark-red flowers, representing the only natural evolution of red flowers in the *Phlox* clade[12]. Both species of *Phlox* and colours of *P. drummondii* are pollinated by the same array of species of Lepidoptera[12]. Hybrids between these two species are formed at rates as high as 11% in the area of sympatry[12]. The hybrids are vigorous but have high, although not complete, male and female sterility[13–15]. Experimental crosses indicate as many as 40% of hand-pollinated hybrid flowers will mature at least one seed and as many as 72% of crosses sired by hybrid pollen will set one seed[15]. Additionally, allozyme data show low levels of gene flow between these species of *Phlox*[14]. Although other traits may contribute to prezygotic isolation (including possible reinforcement traits such as self-compatibility[12]), Levin demonstrated that the shift from light-blue to dark-red flowers in *P. drummondii* decreases interspecific hybridization by 66%, indicating that the change in flower colour substantially increases prezygotic reproductive isolation[12]. Given the above estimates, the hybridization rate before the evolution of dark-red flower colour could have been as high as 28%, which, with low hybrid fitness, would presumably create strong selection to decrease hybridization.

We determined that the evolutionary transition from light-blue to dark-red flower colour in *P. drummondii* results from changes of large effect at two loci. F_2 populations derived from crosses between the allopatric colour variant (light blue) and the sympatric colour variant (dark red) segregate four discrete flower colours: dark blue, light blue, dark red and light red (Fig. 1a). Quantification of the spectral reflectance of 200 F_2 flowers, transformed into two-dimensional Commission internationale de l'éclairage (CIE) 1976 colour space[16], followed by discriminant analysis verified our discrete classifications (Supplementary Fig. 1 and Supplementary Table 4a, b). The ratios of counts within these categories are very close to the 9:3:3:1 ratios expected from two loci with complete dominance ($\chi^2_{(3,\ N=618)} = 0.92$, $P = 0.8206$) (Supplementary Table 3). One locus, *H*, determines flower hue, with blue allele dominant to red, whereas the second locus, *I*, determines colour intensity, with the dark allele dominant to the light.

These two loci appear to determine the types and amounts of anthocyanin floral pigment produced in *P. drummondii*. Anthocyanin pigments, the final products of the well-characterized and highly conserved ABP, are derived from six common types of anthocyanidin by the addition of sugar and/or methyl moieties[17]. Less-hydroxylated anthocyanidins give rise to redder pigments whereas more-hydroxylated anthocyanidins give rise to bluer pigments[17]. Correspondingly, blue-flowered *P. drummondii* (*H-*) produce anthocyanins derived from both the less-hydroxylated cyanidin and peonidin pigments, as well as the more-hydroxylated malvidin pigment, whereas red flowers (*hh*) produce exclusively the less-hydroxylated pigments (Fig. 1b). The change in floral hue thus results from redirecting flux from the malvidin branch of the anthocyanin pathway to the cyanidin/peonidin branch (Fig. 1c). Individuals with increased colour intensity (*I-*) produce more pigment than *ii* individuals, without an effect on pigment composition (Fig. 1b and Supplementary Table 5a, b). These biochemical patterns, coupled with the structure of the ABP (Fig. 1c), suggest candidate genes for the hue and intensity loci.

The structure of the ABP (Fig. 1c) suggests that the loss of flux down the malvidin branch of the ABP might result from changes in one of three candidate genes: (1) loss of function or reduced expression of the gene coding for the branching enzyme flavanoid 3′5′-hydroxylase (F3′5′H); (2) alteration of the substrate specificity owing to a coding mutation in the gene for dihydroflavonol 4-reductase (DFR) making the enzyme unable to metabolize the malvidin precursor dihydromyricitin; and (3) a similar alteration of the substrate specificity of anthocyanidin synthase (ANS). To examine these possibilities, we analysed multiple F_2 populations for co-segregation of floral hue with single nucleotide polymorphism (SNP) markers in the candidate genes. In a total of 100

[1]Department of Biology, Box 90338, Duke University, Durham, North Carolina 27708, USA.

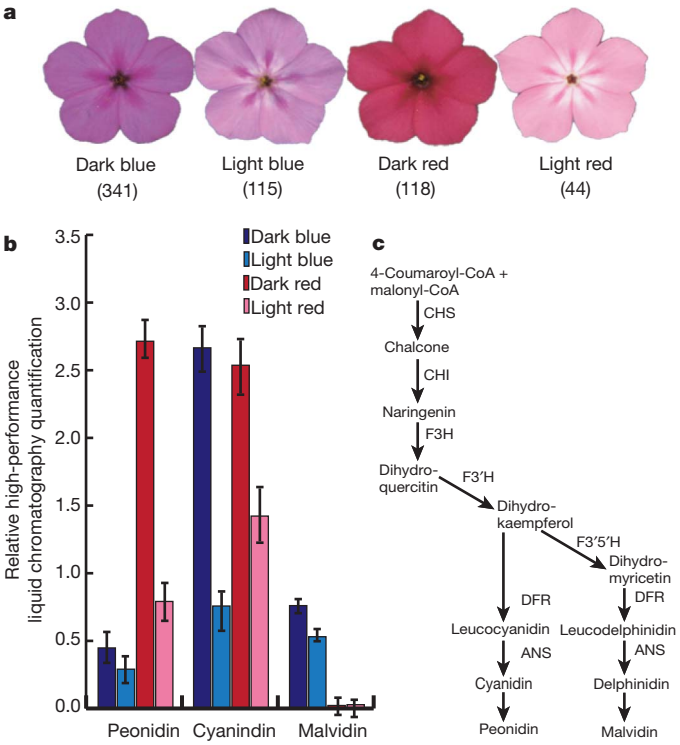


Figure 1 | Flower colour phenotypes in F_2 individuals. **a**, Representative pictures of the four flower colours in F_2 populations: dark blue, light blue, dark red, light red (from left to right). Total counts in F_2 populations indicated under each flower. Ratios of counts are similar to 9:3:3:1 ($\chi^2_{(3,\ N=618)} = 0.92$, $P = 0.8206$). **b**, Relative anthocyanidin pigment production results in variation in F_2 flower colour. Production of all three pigments is significantly different between hue classes. The light- and dark-red flowers produce no malvidin ($F_{(1,25)} = 134.03$, $P < 0.0001$), and significantly more peonidin ($F_{(1,25)} = 58.88$, $P < 0.0001$) and cyanidin ($F_{(1,25)} = 21.57$, $P = 0.0002$). The amount of pigment production is significantly different between classes of flower intensity, with high-intensity flowers producing more of each flower colour pigment ($F_{\text{malvidin}\ (1,25)} = 8.64$, $P = 0.0092$; $F_{\text{peonidin}\ (1,25)} = 32.11$, $P < 0.0001$; $F_{\text{cyanidin}\ (1,25)} = 66.74$, $P < 0.0001$). Standard errors are shown. See Supplementary Table 5a–c for full data and multivariate analysis of variance (MANOVA) results. **c**, A simplified schematic of the ABP, showing core enzymes to the right of the arrows, with the substrates and products indicated at the ends of the arrows. The two branches of the pathway active in *P. drummundii* produce two 3′-hydroxylated red pigments (cyanidin and peonidin) and one 5′-hydroxylated blue pigment (malvidin).

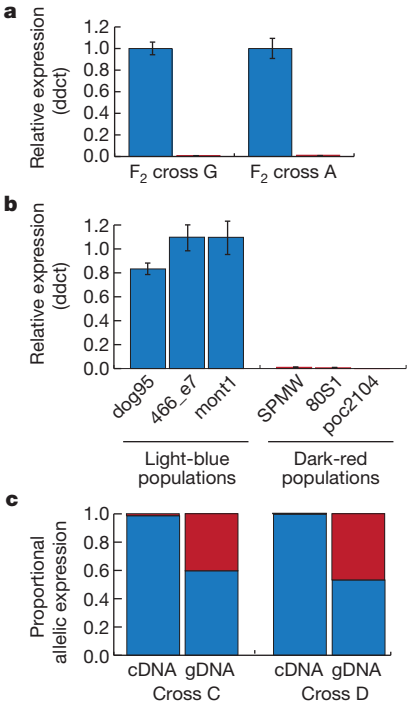


Figure 2 | Results of expression experiments on the hue locus (*F3′5′h*). Quantitative PCR results showing relative expression of *F3′5′h* in blue and red individuals in two F_2 populations (**a**) and six field populations (**b**). Transcripts of *F3′5′h* in red floral tissue (shown in red) are significantly lower than levels detected in blue tissue (shown in blue) in both F_2 populations ($F_{(1,26)} = 250.89$, $P < 0.0001$) and field-collected individuals ($F_{(1,36)} = 30.21$, $P = 0.0053$). Delta–delta cycle-threshold (ddct) indicates relative transcript levels, and bars indicate two standard error units. **c**, Allele-specific expression of *F3′5′h* in multiple heterozygous individuals (*Hh*) from two crosses. Each bar represents the relative contribution of the red allele (shown in red) and the blue allele (shown in blue) to the total expression detected in heterozygous individuals from F_2 families C and D. Relative allelic representation in complementary DNA (cDNA) from heterozygous individuals is significantly different from the relative allelic representation in genomic DNA (gDNA) ($F_{(1,32)} = 375.93$, $P < 0.0001$), indicating a *cis*-regulatory change controlling expression of the red allele. The genomic samples from heterozygous individuals show nearly equal allelic representation (0.5). See Supplementary Fig. 3 for experimental control data.

individuals there was perfect co-segregation between *F3′5′h* and floral hue (Supplementary Table 7). Moreover, genotype at *F3′5′h* explains 77% of the variation in flower hue.

This genetic association corresponds to a downregulation of *F3′5′h* in red flowers. Among F_2 individuals, there is a nearly 100-fold decrease in *F3′5′h* transcripts in red-flowered compared with blue-flowered individuals (Fig. 2a). In addition, red- and blue-flowered individuals collected from multiple populations throughout the range of *P. drummondii* have a comparable difference in *F3′5′h* expression (Fig. 2b) These results demonstrate three features of the genetic change associated with the shift to red flowers: (1) variation in hue is associated with transcript level of *F3′5′h*; (2) genotype at *F3′5′h* predicts transcript level; and (3) the expression difference evident between naturally occurring flower colour variants segregates as a single locus in F_2 individuals.

These patterns imply that expression variation at *F3′5′h* is caused by variation in a *cis*-regulatory element. Allele-specific expression assays confirm this inference: in heterozygous (*Hh*) individuals, the red allele is almost completely downregulated (Fig. 2c). This allelic imbalance indicates a *cis*-regulatory change, whereas equal expression of the two alleles would indicate *trans*-acting regulation of expression[18].

Based on the structure of the ABP, we identified three candidate genetic changes that could explain the increased pigment production in dark (*I*-) flowers: (1) a *cis*-regulatory change that increases production of an ABP enzyme with control over pathway flux; (2) increased enzymatic efficiency (through coding-sequence mutations) of a rate-controlling enzyme in the ABP; and (3) increased expression or function of an ABP transcription factor coordinately regulating the expression of several enzymes. To evaluate these possibilities, we cloned and sequenced genes coding for the core enzymes of the pathway (Fig. 1c), as well as an R2R3-Myb transcription factor orthologous to those known to regulate ABP enzymes in other angiosperms. Of these genes, the marker in *R2R3-Myb* co-segregates perfectly with flower colour intensity in 100 F_2 individuals (Supplementary Table 8) and genotype explains 71% of the intensity variation. These observations suggest *Myb* corresponds to the intensity locus.

This correspondence is further supported by quantification of *Myb* transcript in F_2 individuals. *Myb* expression is significantly higher in dark individuals than light individuals (Fig. 3a), demonstrating that both genotype and expression of this transcription factor are associated with variation in pigment production. Furthermore, field-collected individuals show the same association between expression of *Myb* and flower colour intensity (Fig. 3b). Finally, analysis of allele-specific expression in heterozygous (*Ii*) individuals show that the 'dark' allele is

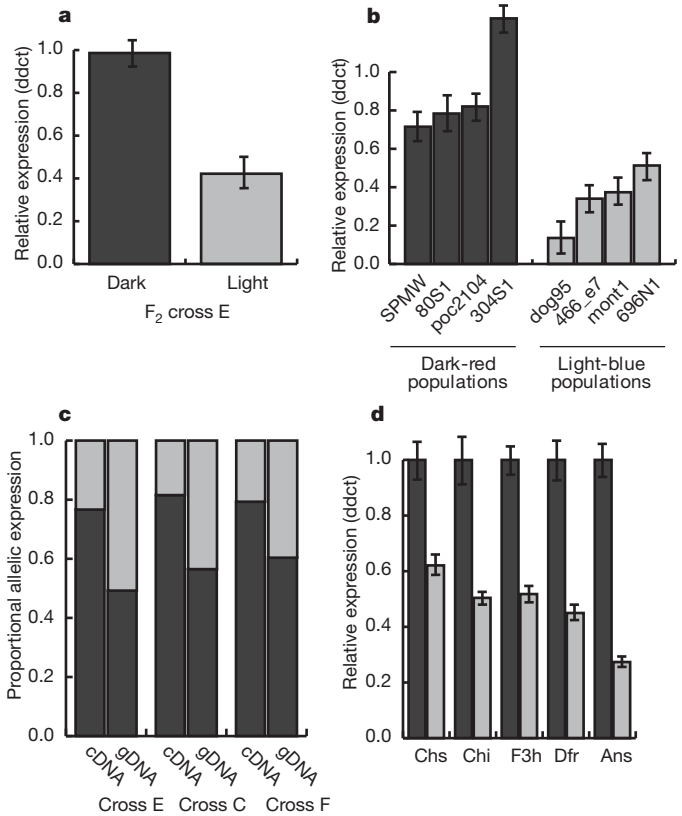


Figure 3 | Results of expression experiments of the intensity locus (*R2R3-Myb*). Relative expression of *R2R3-Myb* in dark and light F_2 individuals (**a**) and natural field populations (**b**). There is a significant upregulation associated with dark individuals relative to light both in F_2 individuals ($F_{(1,12)} = 8.11$, $P = 0.0173$) and in field-collected individuals ($F_{(1,50} = 14.91$, $P = 0.0023$). **c**, Allele-specific expression indicates significantly different allelic representation in cDNA relative to gDNA in heterozygous individuals (*Ii*) from three F_2 families ($F_{(1,12)} = 116.74$, $P < 0.0001$). The over-representation of the dark allele (dark grey) relative to the light allele (light grey) indicates a *cis*-regulatory change. See Supplementary Fig. 4 for complete control data. **d**, There is significant upregulation of transcript levels of all core ABP enzymes in field-collected dark individuals (dark grey) relative to light individuals (light grey). $F_{chs\ (1,36)} = 5.19$, $P = 0.0305$; $F_{chi\ (1,36)} = 5.07$, $P = 0.032$; $F_{f3h\ (1,36)} = 15.81$, $P = 0.0004$; $F_{dfr\ (1,36)} = 13.45$, $P = 0.001$; $F_{ans\ (1,36)} = 36.7$, $P < 0.0001$. For full MANOVA results see Supplementary Table 10. Standard errors are indicated on graphs.

upregulated relative to the 'light' allele (Fig. 3c), a pattern indicative of a *cis*-regulatory change at the *Myb* locus.

In all species that have been examined, this *Myb* coordinately activates several, if not all, of the ABP enzyme-coding genes[19]. Thus we expect that if changes in expression of *Myb* influence pigment intensity, there should be correlated expression changes in the ABP enzyme-coding genes. This expectation was realized: all five core-enzyme genes exhibited significant upregulation in dark flowers relative to light flowers (Fig. 3d). These results indicate that a *cis*-regulatory mutation at the *R2R3-MYB* transcription factor causes the increased colour intensity of *P. drummondii* flowers in sympatric populations.

Our investigation has provided the first identification of genetic changes causing reinforcement of species boundaries. Ideally, we would like to have confirmed the identity of the hue and intensity loci by either fine mapping or transformation, but this is not currently feasible in non-model systems like *P. drummondii*. Nevertheless, we are confident that we have identified the correct genes. Not only do *F3′5′h* and *Myb* co-segregate perfectly with the hue and intensity loci, respectively, but both exhibit changes in expression levels in the directions that are expected to produce the observed phenotypic changes. Moreover, the alternative possibility that the changes are due to linked transcription factors is ruled out by differences in allele-specific expression.

We have shown that reinforcement may involve changes in a few genes, each change having a large phenotypic effect. Our results expand upon two previous analyses of the genetic architecture of reinforcement, which report the involvement of a small number of quantitative trait loci[10,11]. This simple genetic architecture is consistent with theoretical expectations, which indicate that selection for reinforcement is most likely to result in increased reproductive isolation when the phenotypic effect of the assortative mating allele is large and selection for reproductive isolation is strong[7,20].

Ortiz-Barrientos *et al.*[10] suggest that reinforcing reproductive isolation should be inherited as a dominant trait because of Haldane's sieve (the greater probability of a new dominant adaptive mutation reaching fixation than a recessive mutation[21]). The intensity locus fits this expectation, with the derived dark allele dominant to the light allele. In contrast, the hue locus shows the reverse pattern, with the derived red allele recessive to the ancestral blue allele. Although it is easier for selection to increase the frequency of a novel dominant allele, Haldane's sieve can be overcome with strong selection[21]. Additionally, probability of fixation has been found to be independent of dominance when adaptations are not new mutations but are standing genetic variation[22]. Either of these possibilities could explain the fixation of the red allele in sympatry.

Much work investigating the process of speciation focuses on categorizing traits as a particular type of reproductive isolation mechanism (that is, prezygotic or postzygotic) (reviewed by Lowry *et al.*[23] and Nosil *et al.*[24]). Recently it has become clear that a single trait can affect multiple types of reproductive isolation[25,26], but it remains unclear how commonly this occurs. Although the most obvious effect of floral-colour evolution in *P. drummondii* is increased premating isolation with *P. cuspidata*, the associated genetic changes may also have led to increased postzygotic isolation. Hybrids between the dark-red-flowered *P. drummondii* and the light-blue-flowered *P. cuspidata* have dark-blue flowers, which differ from the two parental species. Levin has shown that pollinators discriminate against this intermediate hybrid flower colour[27]. Although further experimental tests are required, this discrimination probably reduces male outcross success and possibly seed set, causing reduced fitness of the hybrids beyond that associated with intrinsic postzygotic isolation. Our work shows how the pattern of genetic dominance can influence whether a trait can contribute to multiple types of isolation. This hypothesized extrinsic postzygotic isolation arises only because the novel allele at the intensity locus is dominant, whereas the novel allele at the hue locus is recessive. Had the novel alleles at both loci exhibited the same dominance, hybrids would have had the same flower colour as one of the parents.

Finally, both of the genetic changes contributing to reinforcement in *P. drummondii* involve *cis*-regulatory mutations, rather than functional (coding-sequence) mutations in pathway enzymes. The types of mutation described here are similar to those described for other cases of flower colour evolution in which reinforcement is not known to be involved[28–30]. These similarities suggest that the genetic basis of reinforcement may be comparable to that of adaptations not associated with reinforcement. However, only by indentifying the genetic basis of reinforcement in other systems can we understand what aspects of the molecular genetic basis of these traits are generally important for reinforcement and which are unique to a particular system.

METHODS SUMMARY

F_2 populations were created from crossing parents collected from allopatry with those collected from sympatry (Supplementary Table 1). F_2 individuals were phenotyped for flower colour and sampled for DNA and RNA extractions. Anthocyanidins were extracted from floral tissue using standard protocol and identified and quantified using high-performance liquid chromatography[30]. Candidate genes were amplified from parental individuals to identify SNPs, and F_2 individuals were genotyped. The Roche Universal ProbeLibrary system was used to quantify expression of all candidate

genes from F_2 and field-collected populations. Pyrosequencing technology was used to quantify allele-specific expression in individuals heterozygous at flower colour loci[18].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 23 August; accepted 2 November 2010.
Published online 9 January 2011.

1. Noor, M. A. F. & Feder, J. L. Speciation genetics: evolving approaches. *Nature Rev. Genet.* **7,** 851–861 (2006).
2. Wolf, J. B. W., Lindell, J. & Backstrom, N. Speciation genetics: current status and evolving approaches. *Phil. Trans. R. Soc. B* **365,** 1717–1733 (2010).
3. Dobzhansky, T. *Genetics and the Origin of Species* (Columbia University Press, 1937).
4. Howard, D. J. in *Hybrid Zones and the Evolutionary Process* (ed. Harrison, R. G.) 46–69 (Oxford University Press, 1993).
5. Servedio, M. R. & Noor, M. A. F. The role of reinforcement in speciation: theory and data. *Annu. Rev. Ecol. Evol. Syst.* **34,** 339–364 (2003).
6. Butlin, R. Speciation by reinforcement. *Trends Ecol. Evol.* **2,** 8–13 (1987).
7. Felsenstein, J. Skepticism towards Santa Rosalia, or why are there so few kinds of animals. *Evolution* **35,** 124–138 (1981).
8. Ortiz-Barrientos, D., Grealy, A. & Nosil, P. The genetics and ecology of reinforcement implications for the evolution of prezygotic isolation in sympatry and beyond. *Year Evol. Biol.* **2009,** 156–182 (2009).
9. Pfennig, K. S. & Pfennig, D. W. Character displacement: ecological and reproductive responses to a common evolutionary problem. *Q. Rev. Biol.* **84,** 253–276 (2009).
10. Ortiz-Barrientos, D., Counterman, B. A. & Noor, M. A. F. The genetics of speciation by reinforcement. *PLoS Biol.* **2,** 2256–2263 (2004).
11. Saether, S. A. *et al.* Sex chromosome-linked species recognition and evolution of reproductive isolation in flycatchers. *Science* **318,** 95–97 (2007).
12. Levin, D. A. Reproductive character displacement in *Phlox*. *Evolution* **39,** 1275–1281 (1985).
13. Levin, D. A. Hybridization between annual species of *Phlox* – population structure. *Am. J. Bot.* 54**,** 1122-& (1967).
14. Levin, D. A. Interspecific hybridization, heterozygosity and gene exchange in *Phlox*. *Evolution* **29,** 37–51 (1975).
15. Ruane, L. G. & Donohue, K. Pollen competition and environmental effects on hybridization dynamics between *Phlox drummondii* and *Phlox cuspidata*. *Evol. Ecol.* **22,** 229–241 (2008).
16. Gonnet, J. F. CIELab measurement, a precise communication in flower color: an example with carnation (*Dianthus Caryophyllus*) cultivars. *J. Hortic. Sci.* **68,** 499–510 (1993).
17. Holton, T. A. & Cornish, E. C. Genetics and biochemistry of anthocyanin biosynthesis. *Plant Cell* **7,** 1071–1083 (1995).
18. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Evolutionary changes in *cis* and *trans* gene regulation. *Nature* **430,** 85–88 (2004).
19. Koes, R., Verweij, W. & Quattrocchio, F. Flavonoids: a colorful model for the regulation and evolution of biochemical pathways. *Trends Plant Sci.* **10,** 236–242 (2005).
20. Kirkpatrick, M. Reinforcement and divergence under assortative mating. *Proc. R. Soc. Lond. B* **267,** 1649–1655 (2000).
21. Haldane, J. B. S. A mathematical theory of natural and artificial selection, part 1. *Trans. Cambr. Phil. Soc.* **23,** 19–41 (1924).
22. Orr, H. A. & Betancourt, A. J. Haldane's sieve and adaptation from the standing genetic variation. *Genetics* **157,** 875–884 (2001).
23. Lowry, D. B., Modliszewski, J. L., Wright, K. M., Wu, C. A. & Willis, J. H. The strength and genetic basis of reproductive isolating barriers in flowering plants. *Phil. Trans. R. Soc. B* **363,** 3009–3021 (2008).
24. Nosil, P., Vines, T. H. & Funk, D. J. Perspective: reproductive isolation caused by natural selection against immigrants from divergent habitats. *Evolution* **59,** 705–719 (2005).
25. Dambroski, H. R. *et al.* The genetic basis for fruit odor discrimination in *Rhagoletis* flies and its significance for sympatric host shifts. *Evolution* **59,** 1953–1964 (2005).
26. Lowry, D. B. & Willis, J. H. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol.* **8,** e1000500 (2010).
27. Levin, D. A. The exploitation of pollinators by species and hybrids of *Phlox*. *Evolution* **24,** 367–377 (1970).
28. Hoballah, M. E. *et al.* Single gene-mediated shift in pollinator attraction in *Petunia*. *Plant Cell* **19,** 779–790 (2007).
29. Schwinn, K. *et al.* A small family of MYB-regulatory genes controls floral pigmentation intensity and patterning in the genus *Antirrhinum*. *Plant Cell* **18,** 831–851 (2006).
30. Des Marais, D. L. & Rausher, M. D. Parallel evolution at multiple levels in the origin of hummingbird pollinated flowers in *Ipomoea*. *Evolution* **64,** 2044–2054 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Des Marais, J. Tung, S. Johnsen and T. Juenger for technical advice, D. Levin for assistance in locating natural populations, and M. Noor for comments on the manuscript. This work was supported by a National Science Foundation grant to M.D.R. and a National Science Foundation Doctoral Dissertation Research Improvement Grant to R.H. R.H. was supported in part by the National Science Foundation Graduate Research Fellowship Program.

Author Contributions R.H. and M.D.R. designed the project; R.H. performed the experiments and the analyses; R.H. and M.D.R. wrote the paper.

Author Information The DNA sequences reported here are deposited in GenBank under accession numbers HQ127319–HQ127344 and HQ323688–HQ323691. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to R.H. (robin.hopkins@duke.edu) or M.D.R. (mrausher@duke.edu).

METHODS

F_2 crosses. Seeds were collected from plants in six populations (Supplementary Table 1) in the spring of 2006, and germinated and grown at the Duke University greenhouses. Seeds were soaked for 48 h in 500 p.p.m. giberellic acid, planted in Metro-Mix 360 (Sun Gro Horticulture), and stratified for 6 days at 4 °C. Plants were transplanted at the four-true-leaf stage into a 50:50 mix of Fafard 4P (Conrad Fafard) and PermaTill One Time (Carolina Stalite Company). Light-blue plants were crossed with dark-red plants to form F_1 seeds. F_1 seeds were grown as above and self-fertilized to create F_2 populations. Self-incompatibility was overcome using bud pollination.

Flower colour phenotyping. All F_2 individuals were categorically phenotyped for flower colour, and a χ^2 test was used to determine if the ratio of categorical flower colour counts differed from 9:3:3:1 (Supplementary Table 3). A subset of individuals were quantitatively phenotyped with a StellarNet EPP2000C spectrometer with a SL1 visible light source and a SL5 deuterium + halogen ultraviolet–visible light source (StellarNet). Raw reflectance was transformed into two-dimensional CIE 1976 LUV colour space[16] (Supplementary Fig. 1). This transformation creates two axes of variation: the x axis corresponds to u' and the y axis corresponds to v'. The white point is located at ($v' = 0.4683305$, $u' = 0.197833$). The v' coordinate represents a measure of hue, and distance from the white point is a quantitative measure of intensity. We used canonical discriminant analysis to confirm that flower colour grouped into four discrete categories (Supplementary Table 4a, b). Anthocyanindins were extracted from a total of 26 individuals (12 from cross A, 14 from cross B) using standard methods of dissolving petal tissue in HCl followed by isoamyl alcohol extractions[30]. Pigments were identified using high-performance liquid chromatography as described in Des Marais and Rausher[30]. Pigments were quantified by calculating the area under the high-performance liquid chromatography peak and scaled to the corresponding standards. We used a MANOVA to determine if pigment amount differed by intensity, hue, family and all interactions of main effects. We used subsequent individual ANOVA models to determine significance of each of the above effects on amount of individual pigments (Supplementary Table 5a, b). All analyses used SAS software version 9.1 (SAS Institute).

Genetic association. Leaf tissue was collected from F_2 and parental individuals. A modified cetyltrimethylammonium bromide (CTAB) extraction was used to isolate the DNA (as in Kelly and Willis[31] but with an additional 2% Triton X-100 added to the CTAB solution and a 3 M sodium acetate wash after ethanol precipitation). Genes in the ABP were amplified, first using degenerate primers designed from orthologous genes in closely related species and then with species-specific primers in subsequent amplifications (Supplementary Table 6). Parental sequences for each gene were used to identify SNPs segregating in the F_2 populations. A subset of F_2 individuals were genotyped at each ABP gene. For those genes that showed an association, subsequent F_2 individuals were genotyped. χ^2 tests were used to confirm associations between genotype and phenotype (Supplementary Tables 7 and 8). A mixed-model ANOVA with F_2 family as a random effect was used to determine how much quantitative flower colour variation was explained by *F3′5′h* and the *R2R3-Myb* genotype. For this analysis we used flower colour reflectance transformed into CIE colour space to determine quantitative measures of hue and intensity (see above). The y axis, corresponding to the value of v', is the measure of hue and the distance each flower colour point is from the white point ($v' = 0.4683305$, $u' = 0.197833$) is the measure of intensity. All analyses used SAS software version 9.1 (SAS Institute).

RNA expression analyses. All expression analyses were performed on flower-bud tissue collected approximately 2 days before opening. RNA was extracted from individuals using the SpectrumTM Plant Total RNA Kit (Sigma-Aldrich). The Roche Applied Science Universal ProbeLibrary (Roche Diagnostics) was used to quantify expression of each gene in the ABP[32]. Probes sites and primers were designed using the online design centre (http://www.roche-applied-science.com/sis/rtpcr/upl/ezhome.html) (Supplementary Table 9). A Thermo Scientific Verso 1-Step RT-qPCR kit was used to amplify according to the manufacturer's instructions (Fisher Scientific). Twenty-six F_2 individuals were used in the *F3′5′h* expression assay. Twelve F_2 individuals were used in the *Myb* expression assay. Two replicate reactions were performed for each sample and the average cycle-threshold value was used in all analyses. Ef1-α was amplified in each sample to control for total amount of RNA in each extraction. Raw expression data were analysed as in Rieu and Powers[33]. We used a mixed-model ANOVA to detect a significant difference in expression in candidate genes between colour groups. F_2 family was used as a random effect in the model.

We collected seeds from natural populations of *P. drummondii* in both allopatry and sympatry (Supplementary Table 2), grew them in the Duke University greenhouse and extracted RNA from bud tissue as described above. Thirty-six individuals were used from field-collected population expression assays, with an additional 14 individuals from two additional populations for the *Myb* assay. Transcript levels of all genes in the ABP were quantified as above. We used a MANOVA to determine if flower colour intensity has an effect on expression of non-causal ABP genes (*Chs*, *Chi*, *F3h*, *Dfr*, *Ans*). Subsequent ANOVAs were used to determine if each individual gene shows a significant effect of intensity on expression (Supplementary Table 10). All analyses were performed using SAS software version 9.1 (SAS Institute).

Allele-specific expression. RNAs from F_1 individuals and heterozygous F_2 individuals were used to quantify allele-specific expression at both the *F3′5′h* and the *Myb* genes. A schematic of the assay design (Supplementary Fig. 2) shows the SNP identity, as well as amplification and sequencing primer sequences. RNA was extracted as described above from 28 individuals for the *Myb* assays (11 heterozygotes and seven homozygous controls) and from eight individuals for the *F3′5′h* assay (six heterozygotes and two homozygous controls). A reverse transcription reaction was performed using InvitrogenTM SuperScript II (Life Technologies) according to the manufacturer's instructions. DNA was extracted from leaf tissue as described above. For each individual, two replicate DNA and four replicate cDNA PCRs were run. No-template controls and no-sequencing-primer controls were performed as well. Pyrosequencing reactions were run on all samples using PyroMARKTM Q96ID (Qiagen)[18,34]. Experiments on both genes were independently replicated.

31. Kelly, A. J. & Willis, J. H. Polymorphic microsatellite loci in *Mimulus guttatus* and related species. *Mol. Ecol.* **7,** 769–774 (1998).
32. Mouritzen, P. *et al.* The ProbeLibrary®-expression profiling 99% of all human genes using only 90 dual-labeled real-time PCR probes. *Biotechniques* **37,** 492–495 (2004).
33. Rieu, I. & Powers, S. J. Real-time quantitative RT-PCR: design, calculations, and statistics. *Plant Cell* **21,** 1031–1033 (2009).
34. Ahmadian, A. *et al.* Single-nucleotide polymorphism analysis by pyrosequencing. *Anal. Biochem.* **280,** 103–110 (2000).

LETTER

doi:10.1038/nature09637

Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts

Toshiro Sato[1], Johan H. van Es[1], Hugo J. Snippert[1], Daniel E. Stange[1], Robert G. Vries[1], Maaike van den Born[1], Nick Barker[1], Noah F. Shroyer[2], Marc van de Wetering[1] & Hans Clevers[1]

Homeostasis of self-renewing small intestinal crypts results from neutral competition between Lgr5 stem cells, which are small cycling cells located at crypt bottoms[1,2]. Lgr5 stem cells are interspersed between terminally differentiated Paneth cells that are known to produce bactericidal products such as lysozyme and cryptdins/defensins[3]. Single Lgr5-expressing stem cells can be cultured to form long-lived, self-organizing crypt–villus organoids in the absence of non-epithelial niche cells[4]. Here we find a close physical association of Lgr5 stem cells with Paneth cells in mice, both *in vivo* and *in vitro*. CD24$^+$ Paneth cells express EGF, TGF-α, Wnt3 and the Notch ligand Dll4, all essential signals for stem-cell maintenance in culture. Co-culturing of sorted stem cells with Paneth cells markedly improves organoid formation. This Paneth cell requirement can be substituted by a pulse of exogenous Wnt. Genetic removal of Paneth cells *in vivo* results in the concomitant loss of Lgr5 stem cells. In colon crypts, CD24$^+$ cells residing between Lgr5 stem cells may represent the Paneth cell equivalents. We conclude that Lgr5 stem cells compete for essential niche signals provided by a specialized daughter cell, the Paneth cell.

In a Matrigel-based culture system containing EGF, the Wnt agonist R-spondin 1 and the BMP inhibitor noggin[4], single Lgr5 stem cells autonomously grow into crypt-like structures with *de novo* generated stem cells and Paneth cells at their bottom. The remainder of these crypts consists of transit-amplifying cells, which feed into villus-like luminal domains containing post-mitotic enterocytes and goblet cells. Thus, a single Lgr5 intestinal stem cell can generate a continuously expanding, self-organizing organoid reminiscent of normal gut in the absence of a subepithelial cellular niche. Confocal cross-sectioning of crypt bottoms of *Lgr5-EGFP-ires-creERT2* mice revealed an almost geometrical distribution of Paneth cells and Lgr5 stem cells that maximized heterotypic contact area (Paneth–stem cell) and minimized homotypic contact area (Fig. 1a–c). The same intimate contact was observed in the organoid cultures at crypt bottoms (Fig. 1d and Supplementary Movie 1).

The hypothesis that Paneth cells supply essential niche signals was rejected previously[5]. To retest this, stem cells that were sorted from *Lgr5-EGFP-ires-creERT2* mice based on GFP expression[6] were recombined with wild-type Paneth cells sorted for CD24 expression (Fig. 2a, c). Of note, CD24-expressing cells reside between Lgr5 stem cells in colon crypts (Fig. 2b), indicating that these are related to Paneth cells. Indeed, a secretory cell type, distinct from goblet cells, resides at colon crypt bottoms[7]. Stem cells and/or Paneth cells were seeded in round-bottomed plates in 10% Matrigel. Reassociated Lgr5 stem cells typically formed short-lived, cystic clusters (Fig. 2d). Reassociated Paneth cells tended to form larger aggregates (Fig. 2e) that disintegrated after 5 days. In three independent experiments, long-lived GFP organoids were formed in only 6.7 ± 3.3% of 10 wells per experiment containing 500 Lgr5 stem cells each, and in 0% of 10 wells per experiment containing 500 Paneth cells each (we occasionally observed GFP-negative organoids originating from contaminating wild-type Paneth cells). When 500 stem cells and 500 Paneth cells were combined, GFP$^+$ organoids formed in 76.7 ± 8.8% of 10 wells per experiment (Fig. 2f, g). The dynamic reassociation process was illustrated using Lgr5$^+$ cells sorted from a clonal RFP$^+$ organoid culture and Paneth cells from a clonal YFP$^+$ organoid culture. As shown in Supplementary Movie 2 (see also Fig. 2h), multiple cell clusters formed initially. The organoids fused to form one or two large organoids per well (Supplementary Movie 2). Thus, Lgr5 stem cells and Paneth cells appeared to require physical contact. Indeed, Lgr5 stem cells are critically dependent on Notch signals[8–10], which depend on direct cell–cell contact.

We then performed comparative gene expression profiling on stem and Paneth cells. The heat-map in Fig. 3a confirmed the segregation of Paneth cell markers (lysozyme, defensin A1 (ref. 3)) and stem-cell markers (*Lgr5*, *Olfm4*, *Tnfrsf19*, *Cdca7* (ref. 11)). Among the genes

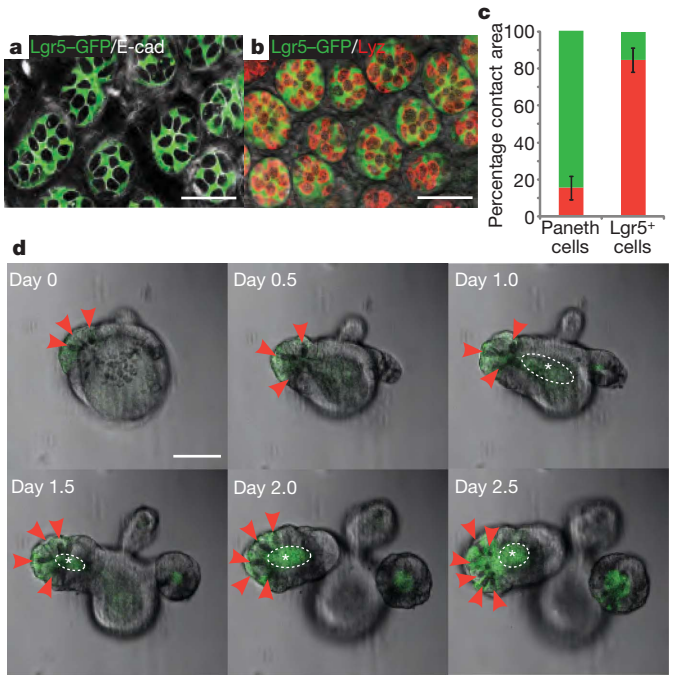


Figure 1 | Geometric distribution pattern of Paneth cells and Lgr5 stem cells. **a**, **b**, Confocal cross-section of *Lgr5-EGFP-ires-creERT2* intestine. E-cadherin (**a**; white) demarcates cell borders. Lgr5 stem cells (green) and Paneth cells (**a**, black; **b**, lysozyme, red) are shown. **c**, Contact area of either Paneth cells or Lgr5 stem cells was quantified with Image J. The values are depicted as mean ± s.d. from three independent mice. Red columns and green columns indicate contact area with Paneth cells and Lgr5 stem cells, respectively. **d**, Stills from Supplementary Movie 1. Time course of crypt organoid growth. Differential interference contrast image reveals granule-containing Paneth cells (red arrowheads) at the site of budding where a new crypt forms. Lgr5–GFP (green) stem cells expand at the crypt base in close proximity to Paneth cells. Asterisk and dotted oval indicate autofluorescence. Scale bar: 50 μm.

[1]Hubrecht Institute, KNAW and University Medical Center Utrecht, Uppsalalaan 8, 3584CT Utrecht, the Netherlands. [2]Cincinnati Children's Hospital, Division of Gastroenterology, Medical Center, MLC 2010, 3333 Burnet Avenue, Cincinnati, Ohio 45229, USA.

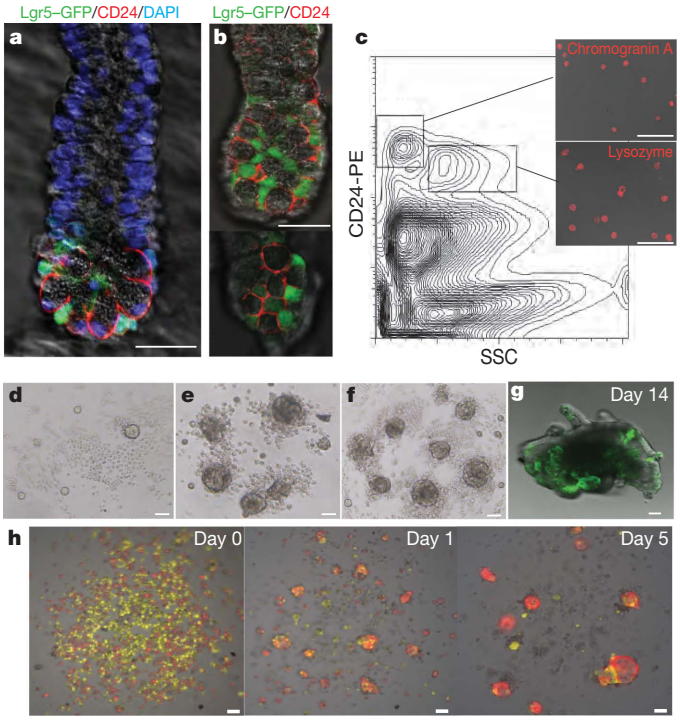


Figure 2 | Paneth cells express CD24 and support growth of Lgr5 stem cells. **a**, Isolated small intestinal crypt. CD24 (red) is expressed by Paneth cells in which granules are visualized by differential interference contrast. Lgr5–GFP^{+} stem cells are green. Counter stain: DAPI (blue). **b**, Isolated colonic crypt. CD24^{+} cells (red) are in intimate contact with Lgr5 stem cells (green). Top, longitudinal crypt section; bottom, section through crypt bottom. **c**, FACS plot of dissociated single cells from small intestinal crypts. Two CD24 bright populations differ by side-scatter (SSC) pattern. Sorted CD24^{hi}SSC^{low} and CD24^{hi}SSC^{hi} cells are subsequently stained. CD24^{hi}SSC^{low} cells are positive for the enteroendocrine marker chromogranin A (top right), whereas CD24^{hi}/SSC^{hi} cells are positive for the Paneth marker lysozyme (bottom right). **d**–**g**, Single sorted Lgr5 stem cells from *Lgr5-EGFP-ires-creERT2* small intestine (**d**), sorted single Paneth cells (**e**) from wild-type small intestine, and a combination of the two cell types (**f**) were seeded in round-bottomed wells and cultured for 2 days. **g**, Lgr5 stem cells form expanding Lgr5–GFP^{+} (green) organoids only when reassociated with Paneth cells. **h**, Stills from Supplementary Movie 2. Time course of the reassociation culture with RFP^{+} Lgr5–GFP stem cells (red) and YFP^{+} Paneth cells (yellow). Scale bar, 50 μm.

most highly enriched in Paneth cells, we noted *Wnt3*, *Wnt11*, *Egf*, *Tgfa* and the Notch ligand *Dll4* (Fig. 3a). *Wnt3* expression[12] and EGF expression[13] had been noted previously. Thus, Paneth cells provided essential signals for stem-cell support: EGF, Wnt3 and Notch. High-level expression of *Wnt3* was confirmed by *in situ* hybridization (Fig. 3b).

R-spondin 1 potently amplifies Wnt responses, yet is inactive on its own[14]. When organoids were grown from crypts derived from Axin2-LacZ mice[15], Wnt responses as assayed by LacZ expression were restricted to the crypt base, despite the ubiquitous presence of R-spondin 1 (Fig. 3c, e and Supplementary Fig. 1). When exogenous Wnt3A was added, the organoids diffusely expressed the blue Wnt reporter (Fig. 4g). The global response to Wnt caused the typical crypt–villus architecture to change into rounded cysts devoid of differentiated cell types (Fig. 3e, g). Indeed, Wnt signalling instructs intestinal cells to adopt a proliferative progenitor phenotype[16]. The same rounded

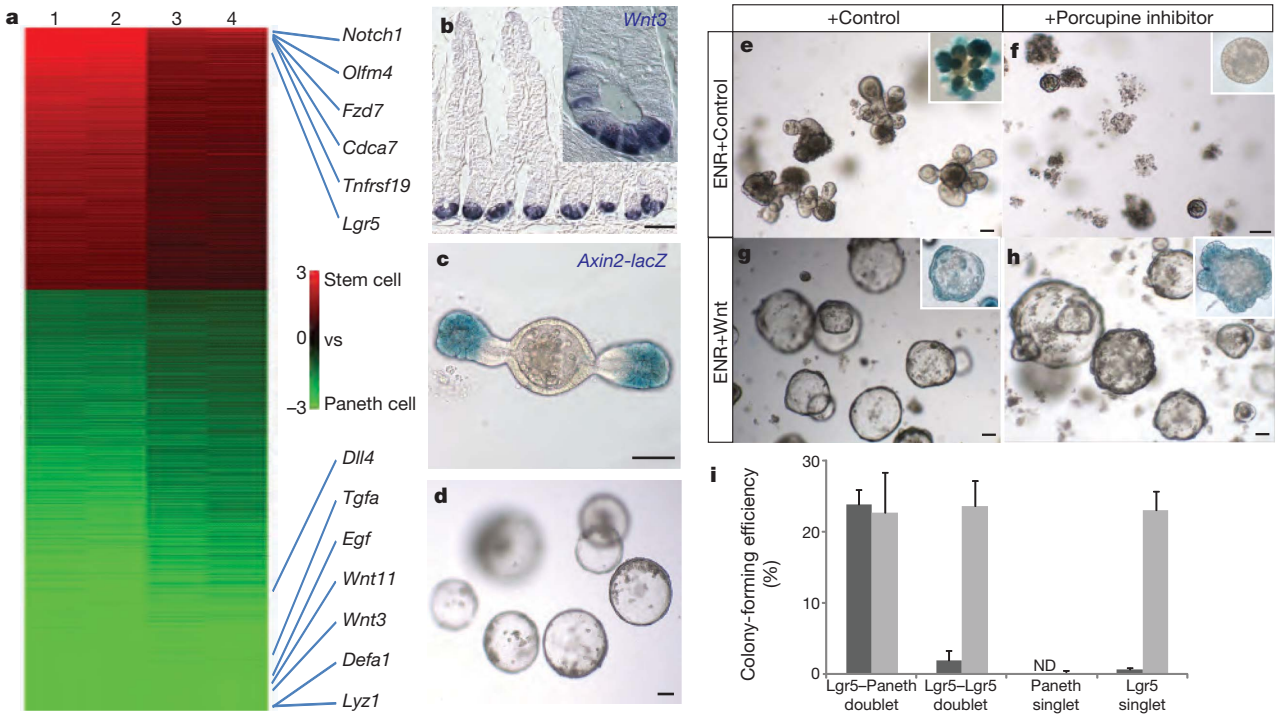


Figure 3 | Paneth cells produce Wnt3 and other essential niche signals for Lgr5 stem cells. **a**, Heat-map of two independent microarray expression experiments (1 and 2; 3 and 4) performed with dye-swap (1 versus 2 and 3 versus 4) from sorted Paneth cells versus Lgr5 stem cells. **b**, *Wnt3* is expressed by Paneth cells at crypt bottoms as analysed by *in situ* hybridization. **c**–**h**, Localized Wnt production regulates crypt–villus morphogenesis in culture. **c**, Freshly isolated crypts from an *Axin2-lacZ* mouse were cultured in standard EGF/noggin/R-spondin 1 medium (ENR) for 4 days. LacZ response is only seen near the bottoms of the two crypts. **d**, Intestinal adenoma samples from *APC^{min}* mice were cultured in ENR medium in the absence of R-spondin for 7 days. **e**, Axin2-LacZ crypts grown in ENR medium. **f**, As in **e**, with the addition of porcupine inhibitor IWP1 at 1 μM. **g**, Crypts from an *Axin2-lacZ* mouse cultured in ENR medium plus Wnt3A. **h**, Same as **g**, with the addition of IWP1. Insets in **e**–**h** depict Axin2-LacZ expression (blue). **e**, **g**, **h**, Six days culture; **f**, 3 days culture after which the organoids disintegrate. See also Supplementary Fig. 2. **i**, Plating efficiency of Lgr5 stem cell–Paneth doublets, Lgr5 stem cell doublets, single Paneth cells and single Lgr5 stem cells with (grey) or without (black) Wnt3A at 100 ng ml^{-1}. Assays were read out as budding organoids at 14 days after sorting. The values are depicted as mean ± standard error of the mean (s.e.m.) from three independent experiments. ND, not detected. See also Supplementary Fig. 3 and Methods for details of doublet isolation and culture. All scale bars, 50 μm.

cysts were routinely observed upon culturing APC-deficient cells from *APC*^{min} adenomas[17] (Fig. 3d). When the small-molecule Wnt secretion inhibitor (porcupine inhibitor) IWP1 (ref. 18) was added, the Axin2-LacZ signal in wild-type organoids was entirely lost and proliferation halted (compare Fig. 3e and f; a dose–response curve is given in Supplementary Fig. 2). This inhibition could be overcome by exogenous Wnt3A (Fig. 3g, h and Supplementary Fig. 2), confirming the specificity of the Wnt secretion inhibitor. We concluded that exogenous R-spondin 1 acts by amplifying the local response to short-range Wnt produced by Paneth cells. Thus, only the direct neighbours of Paneth cells, the Lgr5 stem cells, receive strong Wnt signals, which can be further increased by R-spondin 1. Moreover, these observations indicated that the asymmetry of crypt–villus organoids was established by the localized presence of Wnt-producing Paneth cells. We recently observed that stem cell–Paneth cell doublets display a strongly increased plating efficiency compared to single stem cells[2]. This Paneth-cell-dependence of single stem cells, illustrated in Supplementary Fig. 3, could be overcome by the addition of Wnt3A at 100 ng ml^{-1} for the first 3 days of culture (Fig. 3i).

To investigate, using *in vivo* models, whether Paneth cells provide essential support to Lgr5 stem cells, we used three previously described genetic mouse models for Paneth cell loss: mutation of *Gfi1* (ref. 19), transgenic expression of diphtheria toxin A under the Paneth-cell-specific cryptdin 2 promoter (*CR2-tox176* (ref. 5)) and conditional deletion of *Sox9* (refs 20, 21). When we re-visited crypts of *Gfi1*$^{-/-}$ adult mice, described to lack Paneth cells[19], Paneth cell numbers were reduced but not absent, as also seen recently[22] (Fig. 4a, b and Supplementary Fig. 4). Stem cells were coincidently decreased in number (Fig. 4d, e) and co-localized with remaining Paneth cells, as visualized by double staining for *Olfm4* and lysozyme (Fig. 4g, h). Similarly, in the *CR2-tox176* mice, we noted that Paneth cells were reduced but present (Fig. 4c and Supplementary Fig. 4), in agreement with the reported 82% decrease in Paneth cell numbers[5]. Numbers of stem cells were again decreased, coincident with Paneth cells (Fig. 4c, f, i–k). Lysozyme staining in the sequential intestinal sections from both models revealed that >90% of crypts harboured at least one Paneth cell (not shown). We speculated that the minority of crypts without observable Paneth cells were short-lived, as observed in the conditional *Sox9* model (see below).

We conditionally deleted the *Sox9* gene in 6-week-old mice, homozygous for a *Sox9*^{fl} allele and heterozygous for the *Ah-cre* allele[23]. Paneth cells are estimated to have a lifetime of 8 weeks[24]. The *Sox9* gene was efficiently deleted in all crypt cells with the exception of pre-existing

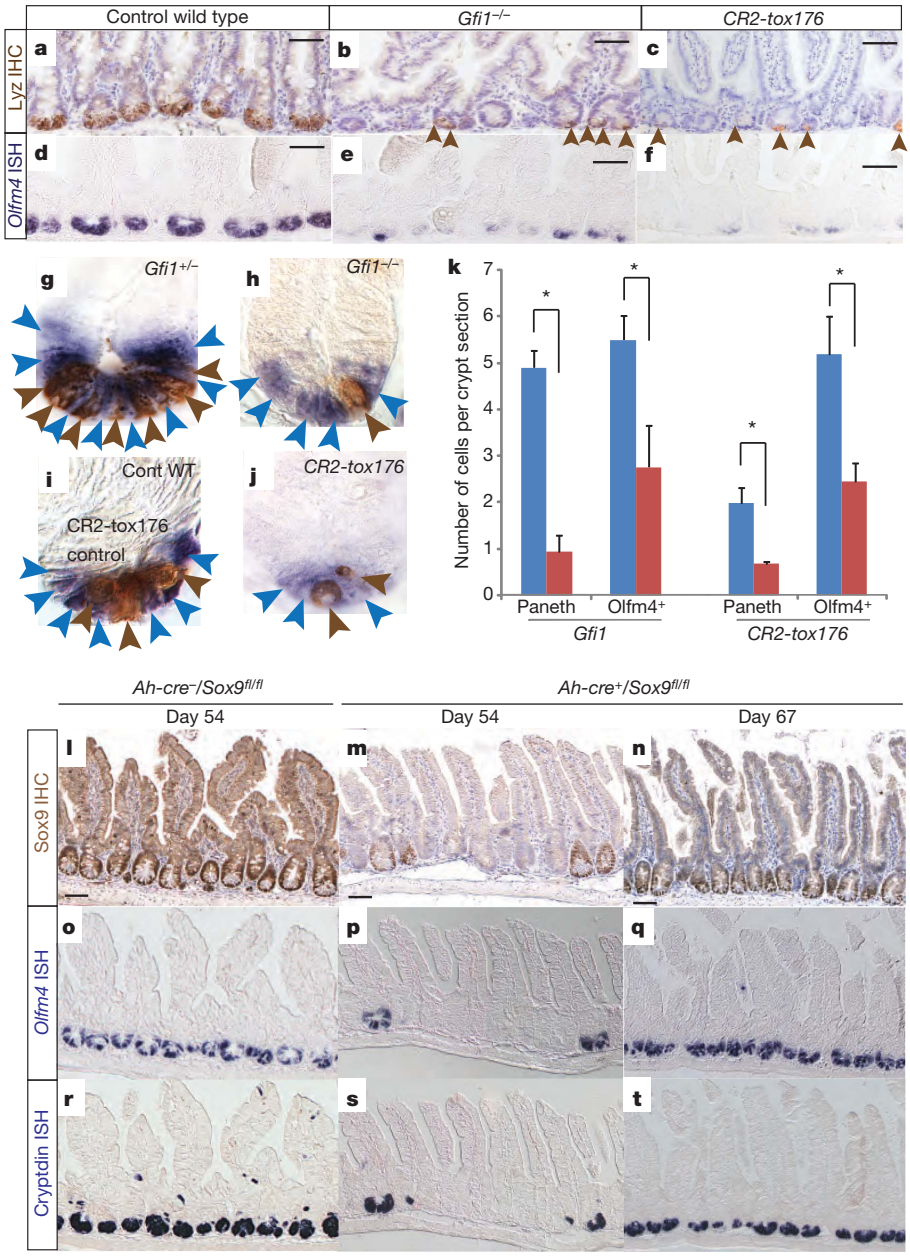


Figure 4 | Paneth cells regulate numbers of intestinal stem cells *in vivo*. **a**–**k**, Paneth cells and stem cells in constitutive models of Paneth cell decrease. **a**–**f**, Lysozyme stain (**a**–**c**; brown arrowheads indicate positive cells) and *Olfm4* staining (**d**–**f**) of crypts of adult (6–7-week-old) mice of the indicated genotype. **g**–**j**, Double stain (lysozyme, brown; *Olfm4*, blue) of a representative crypt of the indicated genotype. Brown and blue arrowheads indicate Paneth cells and stem cells, respectively. **k**, Quantification of stem and Paneth cell numbers in both models. For each bar, 100 crypts were scored for each of three mice. Mutant mice are indicated by red bars and control mice by blue bars. *$P < 0.01$. **l**–**t**, Paneth cells and stem cells in inducible Paneth cell depletion model. Mice of the indicated genotype were injected with β-napthoflavone to induce Cre and were analysed 54 or 67 days after Cre induction by staining for Sox9 protein (brown), *Olfm4* or cryptdin mRNA (blue). Serial sections: **i**, **o**, **r**; **m**, **p**, **s**; and **n**, **q**, **t**. See text for experimental detail. Note the absence of Paneth cells (**s**) and stem cells (**p**) in *Sox9*$^{-/-}$ crypts (**m**). All scale bars, 50 μm.

Paneth cells, where the *Ah-cre* transgene is not activated[23]. Although Sox9 is expressed in Lgr5 stem cells[11], we observed no stem-cell phenotype at early time points after deletion (Supplementary Fig. 5a, b). From 4 weeks onwards, Paneth cell numbers visibly decreased. Loss was virtually complete after 7–8 weeks (Fig. 4s), after which a regenerative response occurred. We occasionally noted $Sox9^{-/-}$ crypts with a single remaining Sox9-positive Paneth cell (Supplementary Fig. 5c; red arrows). Stem cells disappeared coincident with Paneth cells, and the remaining stem cells crowded around the remaining Paneth cells (Supplementary Fig. 5d). Supplementary Fig. 6 depicts a field of 'escaper' wild-type crypts adjacent to a field of Sox9-negative crypts. The escaping wild-type crypts containing abundant Paneth cells rapidly replaced the mutant crypts by crypt fission (Supplementary Fig. 6). By day 67, all crypt basal cells were $Sox9^{+}$ again (Fig. 4n) and contained normal numbers of Paneth cells (Fig. 4t) and stem cells (Fig. 4q). From this, we concluded that Paneth cells are essential for the maintenance of crypts and stem cells.

Stem cell niches are typically portrayed as pre-existing sites to which stem cells migrate[25]. Here we show that intestinal stem cells receive niche support from their own specialized progeny. This is not without precedent, as the somatic stem cells of the fly testis give rise to differentiated cells that in turn build the testis niche[26]. Thus, Paneth cells serve as multifunctional guardians of stem cells, both by secreting bactericidal products and by providing essential niche signals. Lgr5 stem cells divide symmetrically and their numbers are restricted by neutral competition at the stem-cell population level[2]. We now propose that Lgr5 stem cells compete for available Paneth cell surface. Paneth cell numbers must therefore be tightly regulated, which is indeed the case. Paneth cells are generated directly above the crypt base, the latter originally termed the 'stem cell zone'[27,28]. It will be of interest to understand what determines Paneth cell numbers and their slow turnover rate.

METHODS SUMMARY

Reagents. Murine recombinant EGF and noggin were from Peprotech; Wnt3A was from Millipore. Human recombinant R-spondin 1 was provided by A. Abo[29]. Y-27632 was from Sigma; IWP1 was provided by L. Lum[18].

Mice. *Lgr5-EGFP-ires-creERT2* mice[1], APC^{min} (ref. 17), *Axin2-lacZ* mice[15], $Gfi1^{-/-}$ (ref. 19), *CR2-tox17* (ref. 5), $Sox9^{fl/fl}$ (ref. 20) and R26R-confetti[2] mice have been described earlier. The transgenic *Ah-cre* line[23] was crossed with $Sox9^{fl/fl}$ mice. Cre enzyme was induced by intraperitoneal injections of 200 μl β-napthoflavone (10 mg ml^{-1}; Sigma) dissolved in corn oil for three consecutive days.

Crypt isolation, cell dissociation and culture. Crypt isolation, cell dissociation and culture have been described previously[2,4]; see Methods for details.

Reassociation assay. A total of 500 sorted Lgr5–GFP^{hi} stem cells (purity >99%) were co-cultured with 500 genetically unmarked $CD24^{+}$ Paneth cells (purity >95%). Cells were re-suspended in 100 μl of culture medium in Ultra-low attachment 96-well round-bottomed plates (Corning) and the plates were left on ice for 15 min. The plate was centrifuged (300*g*) for 5 min and 10 μl of Matrigel was added to each well. For Supplementary Movie 2, Lgr5–GFP^{hi}/confetti–RFP^{+} and Lgr5–GFP^{hi}/confetti–YFP^{+} stem cells were sorted separately from *Lgr5-EGFP-ires-creERT2* × *R26R*-confetti mice[2], tamoxifen-induced 3 days before they were killed. After 10 days of culture, 1,500 Lgr5–GFP^{hi}/$CD24^{dim}$/confetti–RFP^{+} stem cells and 1,500 $CD24^{hi}$/confetti–YFP^{+} Paneth cells were sorted from these two respective organoid cultures, and filmed for ten consecutive days, interrupted twice for exchange of medium (see Methods for details).

Tissue preparation for confocal analysis. For semi-thick sectioning of near-native tissue, organs were fixed in 4% paraformaldehyde at room temperature for 20 min and washed in cold PBS.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 26 March; accepted 4 November 2010.
Published online 28 November 2010.

1. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene *Lgr5*. *Nature* **449,** 1003–1007 (2007).
2. Snippert, H. J. *et al.* Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143,** 134–144 (2010).
3. Porter, E. M., Bevins, C. L., Ghosh, D. & Ganz, T. The multifaceted Paneth cell. *Cell. Mol. Life Sci.* **59,** 156–170 (2002).
4. Sato, T. *et al.* Single Lgr5 stem cells build crypt–villus structures *in vitro* without a mesenchymal niche. *Nature* **459,** 262–265 (2009).
5. Garabedian, E. M., Roberts, L. J., McNevin, M. S. & Gordon, J. I. Examining the role of Paneth cells in the small intestine by lineage ablation in transgenic mice. *J. Biol. Chem.* **272,** 23729–23740 (1997).
6. Van der Flier, L. G. *et al.* The intestinal Wnt/TCF signature. *Gastroenterology* **132,** 628–632 (2007).
7. Altmann, G. G. Morphological observations on mucus-secreting nongoblet cells in the deep crypts of the rat ascending colon. *Am. J. Anat.* **167,** 95–117 (1983).
8. van Es, J. H. *et al.* Notch/γ-secretase inhibition turns proliferative cells in intestinal crypts and adenomas into goblet cells. *Nature* **435,** 959–963 (2005).
9. Fre, S. *et al.* Notch signals control the fate of immature progenitor cells in the intestine. *Nature* **435,** 964–968 (2005).
10. van Es, J. H., de Geest, N., van de Born, M., Clevers, H. & Hassan, B. A. Intestinal stem cells lacking the Math1 tumour suppressor are refractory to Notch inhibitors. *Nature Communication* **1,** 1–5 (2010).
11. van der Flier, L. G. *et al.* Transcription factor achaete scute-like 2 controls intestinal stem cell fate. *Cell* **136,** 903–912 (2009).
12. Gregorieff, A. *et al.* Expression pattern of Wnt signaling components in the adult intestine. *Gastroenterology* **129,** 626–638 (2005).
13. Poulsen, S. S., Nexo, E., Olsen, P. S., Hess, J. & Kirkegaard, P. Immunohistochemical localization of epidermal growth factor in rat and man. *Histochemistry* **85,** 389–394 (1986).
14. Binnerts, M. E. *et al.* R-Spondin1 regulates Wnt signaling by inhibiting internalization of LRP6. *Proc. Natl Acad. Sci. USA* **104,** 14700–14705 (2007).
15. Lustig, B. *et al.* Negative feedback loop of Wnt signaling through upregulation of conductin/axin2 in colorectal and liver tumors. *Mol. Cell. Biol.* **22,** 1184–1193 (2002).
16. van de Wetering, M. *et al.* The β-catenin/TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells. *Cell* **111,** 241–250 (2002).
17. Moser, A. R., Pitot, H. C. & Dove, W. F. A dominant mutation that predisposes to multiple intestinal neoplasia in the mouse. *Science* **247,** 322–324 (1990).
18. Chen, B. *et al.* Small molecule-mediated disruption of Wnt-dependent signaling in tissue regeneration and cancer. *Nature Chem. Biol.* **5,** 100–107 (2009).
19. Shroyer, N. F., Wallis, D., Venken, K. J., Bellen, H. J. & Zoghbi, H. Y. Gfi1 functions downstream of Math1 to control intestinal secretory cell subtype allocation and differentiation. *Genes Dev.* **19,** 2412–2417 (2005).
20. Mori-Akiyama, Y. *et al.* SOX9 is required for the differentiation of paneth cells in the intestinal epithelium. *Gastroenterology* **133,** 539–546 (2007).
21. Bastide, P. *et al.* Sox9 regulates cell proliferation and is required for Paneth cell differentiation in the intestinal epithelium. *J. Cell Biol.* **178,** 635–648 (2007).
22. Bjerknes, M. & Cheng, H. Cell lineage metastability in Gfi1-deficient mouse intestinal epithelium. *Dev. Biol.* **345,** 49–63 (2010).
23. Ireland, H. *et al.* Inducible Cre-mediated control of gene expression in the murine gastrointestinal tract: effect of loss of β-catenin. *Gastroenterology* **126,** 1236–1246 (2004).
24. Ireland, H., Houghton, C., Howard, L. & Winton, D. J. Cellular inheritance of a Cre-activated reporter gene to determine Paneth cell longevity in the murine small intestine. *Dev. Dyn.* **233,** 1332–1336 (2005).
25. Morrison, S. J. & Spradling, A. C. Stem cells and niches: mechanisms that promote stem cell maintenance throughout life. *Cell* **132,** 598–611 (2008).
26. Voog, J., D'Alterio, C. & Jones, D. L. Multipotent somatic stem cells contribute to the stem cell niche in the *Drosophila* testis. *Nature* **454,** 1132–1136 (2008).
27. Bjerknes, M. & Cheng, H. The stem-cell zone of the small intestinal epithelium. I. Evidence from Paneth cells in the adult mouse. *Am. J. Anat.* **160,** 51–63 (1981).
28. Bjerknes, M. & Cheng, H. The stem-cell zone of the small intestinal epithelium. III. Evidence from columnar, enteroendocrine, and mucous cells in the adult mouse. *Am. J. Anat.* **160,** 77–91 (1981).
29. Kim, K. A. *et al.* R-Spondin proteins: a novel link to β-catenin activation. *Cell Cycle* **5,** 23–26 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank H. Begthel, J. Korving and S. van den Brink for technical assistance; J. Gordon for providing small intestinal sections from *Cr2-tox177* mice; L. Lum for providing IWP1; and A. Abo for R-spondin 1.

Author Contributions T.S. and H.C. conceived and designed the experiments. T.S., J.H.v.E., H.J.S., R.G.V., M.v.d.B., N.B. and M.v.d.W. performed the experiments, and T.S., J.H.v.E., H.J.S., D.E.S. and H.C. analysed the data. N.F.S. provided $Gfi1^{-/-}$ mice intestines. T.S. and H.C. wrote the manuscript.

Author Information The data for the microarray analysis were deposited to the GEO database under accession number GSE25109. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to H.C. (h.clevers@hubrecht.eu).

METHODS

Reagents. Murine recombinant EGF and noggin were from Peprotech. Murine recombinant Wnt3A was from Millipore. Human recombinant R-spondin 1 was provided by A. Abo[29]. Y-27632 was from Sigma. IWP1 was provided by L. Lum[18].

Mice. *Lgr5-EGFP-ires-creERT2* mice[1], *APC*^{min} (ref. 17), *Axin2-lacZ* mice[15], *Gfi1*$^{-/-}$ (ref. 19), *CR2-tox17* (ref. 5), *Sox9*$^{fl/fl}$ (ref. 20) and R26R-confetti[2] mice have been described earlier. The transgenic *Ah-cre* line[23] was crossed with *Sox9*$^{fl/fl}$ mice. Cre enzyme was induced by intraperitoneal injections of 200 µl β-napthoflavone (10 mg ml^{-1}; Sigma Aldrich) dissolved in corn oil for three consecutive days.

Crypt isolation, cell dissociation and culture. Crypt isolation, cell dissociation and culture have been described previously[2,4]. For culture/sorting experiments, at least three independent experiments were performed. For each experiment, crypts/cells were pooled from three intestines. For microarray, sorted cells from ten intestines were pooled. Crypts were directly cultured as previously described (100 crypts per well on 24-well plates)[4]. For single-cell or doublet-cell culture, crypts were dissociated with TrypLE express (Invitrogen) including 2,000 U ml^{-1} DNase (Sigma) for 30 min at 37 °C or 2 h at room temperature. For reassociation assay from established crypt organoids, the samples were dissociated with TrypLE express for 15 min at 37 °C. Dissociated cells were passed through 20-µm cell strainer (Celltrix) and washed with PBS. Cells were stained with PE-conjugated anti-CD24 antibody (eBioscience) and APC-conjugated anti-Epcam antibody (eBioscience) for 15 min at 4 °C, and analysed by MoFlo (DakoCytomation). Viable epithelial single cells or doublets were gated by forward scatter, side scatter and pulse-width parameter, and negative staining for propidium iodide or 7-ADD (eBioscience). Sorted cells were collected, pelleted and embedded in Matrigel (BD bioscience), followed by seeding on a 96-well plate (30–50 singlets or doublets; 10 µl Matrigel per well). Culture medium (Advanced DMEM/F12 supplemented with penicillin/streptomycin, 10 mM HEPES, Glutamax, 1× N2, 1× B27 (all from Invitrogen) and 1 µM *N*-acetylcysteine (Sigma) containing growth factors 50 ng ml^{-1} EGF, 100 ng ml^{-1} noggin, 1 µg ml^{-1} R-spondin) was overlaid. Y-27632 (10 µM) was included for the first 2 days to avoid anoikis. Growth factors were added every other day and the entire medium was changed every 4 days. In some experiments, 100 ng ml^{-1} Wnt3A (Millipore) was added in the culture medium. Sorted cells were manually inspected by inverted microscopy, and the numbers of viable organoids in triplicate were calculated.

Reassociation assay. A total of 500 sorted Lgr5–GFP^{hi} stem cells (purity >99%) were co-cultured with 500 genetically unmarked CD24^{+} Paneth cells (purity >95%). Cells were re-suspended in 100 µl of culture medium in Ultra-low attachment 96-well round-bottomed plates (Corning) and the plate was left on ice for 15 min. The plate was then centrifuged (300*g*) for 5 min and 10 µl of Matrigel was added in each well. For Supplementary Movie 2, Lgr5–GFP^{hi}/confetti–RFP^{+} and Lgr5–GFP^{hi}/confetti–YFP^{+} stem cells were sorted separately from *Lgr5-EGFP-ires-creERT2* × *R26R*-confetti mice[2], tamoxifen-induced three days before the mice were killed. After 10 days of culture, 1,500 Lgr5–GFP^{hi}/CD24^{dim}/confetti–RFP^{+} stem cells and 1,500 CD24^{hi}/confetti–YFP^{+} Paneth cells were sorted from these two respective organoid cultures, and filmed for ten consecutive days, interrupted twice for exchange of medium. The fluorescent and phase-contrast images were acquired every 20 min by inverted microscopy (AF7000, Leica) equipped with a live imaging chamber (humidified with sterile water and maintained at 37 °C, 7.5% CO_2).

Microarray analysis. Single Lgr5–GFP^{hi} cells or Paneth cells were sorted into Buffer RLT in the RNeasy Micro kit (Qiagen). Microarray analysis (Agilent) was performed as previously described[4]. The data were deposited to the GEO database (accession number GSE25109). Heat-maps were created using Treeview software.

Histology, immunohistochemistry and *in situ* hybridization. Samples taken from the middle of the small intestine were fixed in 4% paraformaldehyde (PFA), embedded in paraffin, and processed as previously described[1]. The primary antibodies were: mouse anti-E-cadherin (1:100, BD transduction), mouse anti-Ki67 (1:250, Monosan), mouse anti-phospho histone H3 (1:1,000, Millipore), rabbit anti-Sox9 (1:600, Millipore), rabbit anti-lysozyme (1:1,000, Dako) and anti-chromogranin A (1:100, Santa Cruz). The secondary antibodies were peroxidase-conjugated antibodies or Alexa-568-conjugated antibodies. For whole-mount immunostaining, freshly isolated crypts were fixed with 4% PFA, and stained with anti-CD24 antibody (eBioscience) over night at 4 °C. After washing, the samples were incubated with Alexa-568-conjugated anti-rat antibody over night at 4 °C. DNA was stained by DAPI (Molecular Probe). For counting the number of Lgr5 stem cells (GFP) and Paneth cells (lysozyme), 1 cm^2 of PFA-fixed intestinal wall was put in a mould. Four per cent low-melting-point agarose (40 °C) was added and allowed to cool on ice. Once solid, a vibrating microtome (HM650, Microm) was used to make semi-thick sections (150 µm) (velocity, 1 mm s^{-1}; frequency, 65 Hz; amplitude, 0.9 mm)^2. Sections were permeabilized and stained as previously described[4]. Images were acquired with confocal microscopy (Leica, SP5). For *in situ* hybridization, cRNA probes were generated from full-length cDNA expression vectors (IMAGE consortium or RPZD) with *in vitro* transcription. The protocol was described elsewhere[12]. X-gal staining was performed as previously described[4].

LETTER

doi:10.1038/nature09674

Reduction of disulphide bonds unmasks potent antimicrobial activity of human β-defensin 1

Bjoern O. Schroeder[1,2], Zhihong Wu[3], Sabine Nuding[1,2], Sandra Groscurth[4]†, Moritz Marcinowski[5], Julia Beisner[1,2], Johannes Buchner[5], Martin Schaller[6], Eduard F. Stange[7] & Jan Wehkamp[1,2,7]

Human epithelia are permanently challenged by bacteria and fungi, including commensal and pathogenic microbiota[1,2]. In the gut, the fraction of strict anaerobes increases from proximal to distal, reaching 99% of bacterial species in the colon[3]. At colonic mucosa, oxygen partial pressure is below 25% of airborne oxygen content, moreover microbial metabolism causes reduction to a low redox potential of −200 mV to −300 mV in the colon[4]. Defensins, characterized by three intramolecular disulphide-bridges, are key effector molecules of innate immunity that protect the host from infectious microbes and shape the composition of microbiota at mucosal surfaces[5–8]. Human β-defensin 1 (hBD-1) is one of the most prominent peptides of its class but despite ubiquitous expression by all human epithelia, comparison with other defensins suggested only minor antibiotic killing activity[9,10]. Whereas much is known about the activity of antimicrobial peptides in aerobic environments, data about reducing environments are limited. Herein we show that after reduction of disulphide-bridges hBD-1 becomes a potent antimicrobial peptide against the opportunistic pathogenic fungus *Candida albicans* and against anaerobic, Gram-positive commensals of *Bifidobacterium* and *Lactobacillus* species. Reduced hBD-1 differs structurally from oxidized hBD-1 and free cysteines in the carboxy terminus seem important for the bactericidal effect. *In vitro*, the thioredoxin (TRX) system[11] is able to reduce hBD-1 and TRX co-localizes with reduced hBD-1 in human epithelia. Hence our study indicates that reduced hBD-1 shields the healthy epithelium against colonisation by commensal bacteria and opportunistic fungi. Accordingly, an intimate interplay between redox-regulation and innate immune defence seems crucial for an effective barrier protecting human epithelia.

We modified the radial diffusion assay[12] to analyse antimicrobial activity of synthetic hBD-1 against anaerobic bacteria of the normal flora under anaerobic conditions. Increasing concentrations of the reducing agent dithiothreitol (DTT) were added to assay medium. In medium without DTT hBD-1 did not affect growth of Gram-positive *Bifidobacterium adolescentis* (Fig. 1a). Surprisingly, addition of increasing amounts of DTT led to an increase of inhibition zones in size and sharpness. In contrast, human β-defensin 3 (hBD-3) and lysozyme showed less antimicrobial activity upon addition of DTT (Fig. 1b and Supplementary Fig. 1). Thus, increased antimicrobial activity of hBD-1 in reducing environment was specific for hBD-1 and not caused by impaired bacterial fitness. Remarkably, at concentrations of 2 mM DTT hBD-1 became as effective as hBD-3 against *B. adolescentis* (Fig. 1c). This finding is crucial as hBD-3 is one of the most powerful antimicrobial peptides in oxygen-rich environment, whereas hBD-1 was thought to be one of the weakest.

For the Gram-negative anaerobe *Bacteroides vulgatus* we found no antimicrobial effect of hBD-1 and lysozyme under any conditions, whereas hBD-3 showed concentration-dependent inhibition zones (Supplementary Fig. 2).

Because β-defensins contain three intramolecular disulphide-bridges[13] we investigated the involvement of cystines for the observed antimicrobial effect. Therefore, we incubated hBD-1 with increasing concentrations of DTT and analysed samples using matrix-assisted laser

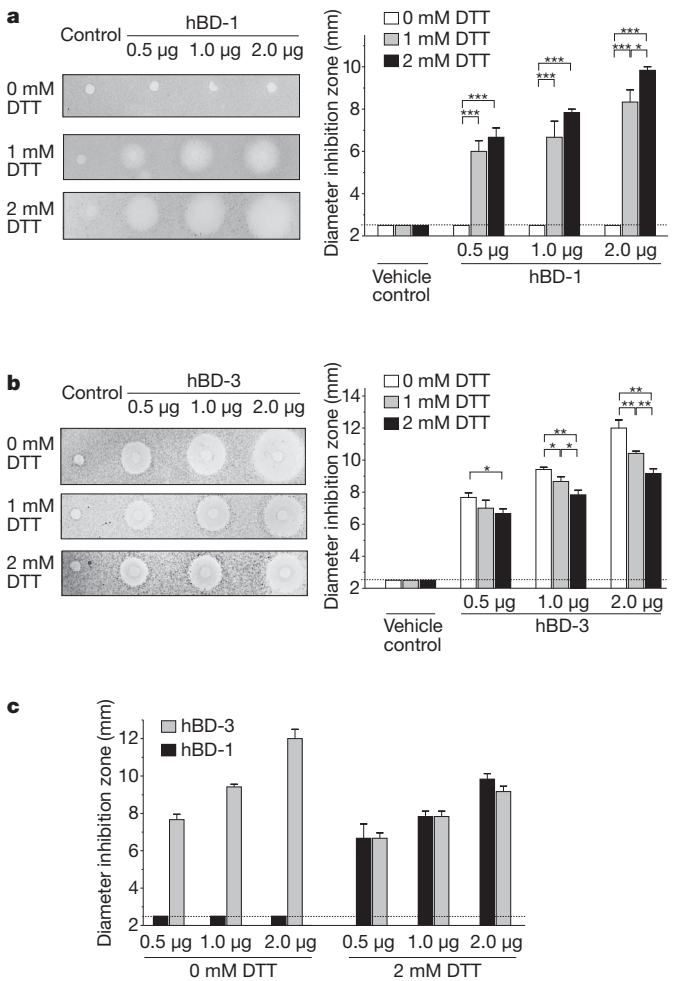


Figure 1 | hBD-1 shows antimicrobial activity under reducing conditions. **a**, **b**, *Bifidobacterium adolescentis* was incubated with up to 2 mM DTT under anaerobic conditions with hBD-1 (**a**) or hBD-3 (**b**). Inhibition zones were measured and statistically evaluated using student's *t*-test with $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. Representative radial diffusion assays of three experiments are shown. Data are presented as means, error bars indicate standard deviation. **c**, Comparison of inhibition zones between hBD-1 and hBD-3 in growth medium without or with 2 mM DTT. Dotted lines at 2.5 mm represent base line diameter of punched wells.

[1]Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, 70376 Stuttgart, Germany. [2]University of Tübingen, 72076 Tübingen, Germany. [3]Department of Dermatology, University Hospital Schleswig-Holstein, Campus Kiel, 24105 Kiel, Germany. [4]Department 1 Protein Evolution, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany. [5]Center for Integrated Protein Science Munich, Department Chemie, Technische Universität München, 85747 Garching, Germany. [6]Department of Dermatology, University Hospital Tübingen, 72076 Tübingen, Germany. [7]Department of Internal Medicine I, Robert Bosch Hospital, 70376 Stuttgart, Germany. †Present address: Bruker BioSpin AG, 8117 Fällanden, Switzerland.

desorption/ionisation (MALDI)-mass spectroscopy (MS). A single signal could be detected for oxidized hBD-1 (oxhBD-1) at *m/z* 3,926.8 (Fig. 2a). DTT-treatment and carboxamidomethylation resulted in a second signal, corresponding to the completely reduced form (redhBD-1) with all six cysteine residues alkylated. These results suggest that all three cystines are either present or absent, without the existence of intermediate states containing one or two disulphide-bridges.

Reversed-phase high performance liquid chromatography (RP-HPLC) showed that oxhBD-1 eluted after 30 min (Fig. 2b). Incubation with increasing concentrations of DTT resulted in a shift towards a peak at 33.5 min, representing the completely reduced hBD-1 (confirmed by MALDI-MS, data not shown) and indicating an increase of hydrophobicity of redhBD-1 in solution.

To elucidate structural changes further, nuclear magnetic resonance (NMR) experiments on recombinant, uniformly ^{15}N-labelled reduced and oxidized hBD-1 were performed. oxhBD-1 showed a well-dispersed ^{15}N-heteronuclear single quantum coherence (HSQC) spectrum as expected for a folded protein (Fig. 2c). In addition, the detected cross peaks in the HNH- and NNH-nuclear Overhauser enhancement spectroscopy (NOESY) spectra as well as the restricted dynamics observed in the ^{1}H-^{15}N-heteronuclear NOE (hetNOE) experiment indicated a highly structured peptide (data not shown). In contrast, the dispersion of the HSQC spectrum of redhBD-1 indicated a lack of hydrogen bonds for most parts of the protein (Fig. 2c). Accordingly, the HNH- and NNH-NOESY spectra showed significantly less cross peaks and the dynamics of the hetNOE experiment clearly support an unstructured, highly flexible polypeptide chain. Similar findings were obtained by circular dichroism (CD) spectroscopy (Fig. 2d). oxhBD-1 displayed a characteristic minimum at 209 nm for alpha-helices and a significant signal at 218 nm, corresponding to the beta-sheet content, indicating a well-folded peptide. In contrast, redhBD-1 did not show these minima but displayed a minimum around 195 nm, indicating a random coil-like structure, being consistent with previous data[14]. Indeed, calculation of secondary structure elements proposed an increase in the random coil fraction for the reduced peptide (Supplementary Table 1). From these structural observations we propose that antimicrobial activity of hBD-1 in a reducing environment is attributable to its reduced, unstructured form.

To exclude artificial effects of the reducing agent DTT we compared the antimicrobial activity of oxhBD-1 and redhBD-1 against different bifidobacteria under anaerobic conditions in medium lacking DTT. Synthetic, oxidized hBD-1 did not affect growth of bifidobacteria whereas linear hBD-1 showed antimicrobial activity against all examined strains (Fig. 3a and Supplementary Fig. 3). Similar results were obtained for Gram-positive lactobacilli, whereas we detected no effect of any hBD-1 against Gram-negative *Bacteroides vulgatus* (Fig. 3b). For the facultative anaerobe *Escherichia coli* K12 we found antimicrobial activity of both oxhBD-1 and redhBD-1, which is in accordance with literature[14–16]. When testing the commensal, facultative pathogenic fungus *Candida albicans* we found potent antifungal activity of redhBD-1 but not oxhBD-1 against four out of five strains (Fig. 3c) when using a flow cytometric antimicrobial killing assay[17]. Consequently, reduced hBD-1 is not only antimicrobial against Gram-positive anaerobes of the human normal flora but also against a facultative pathogenic fungus of clinical importance.

Subsequently, we generated recombinant Ala/Ser-variants of hBD-1 in which all cysteine residues were substituted by alanine or serine to

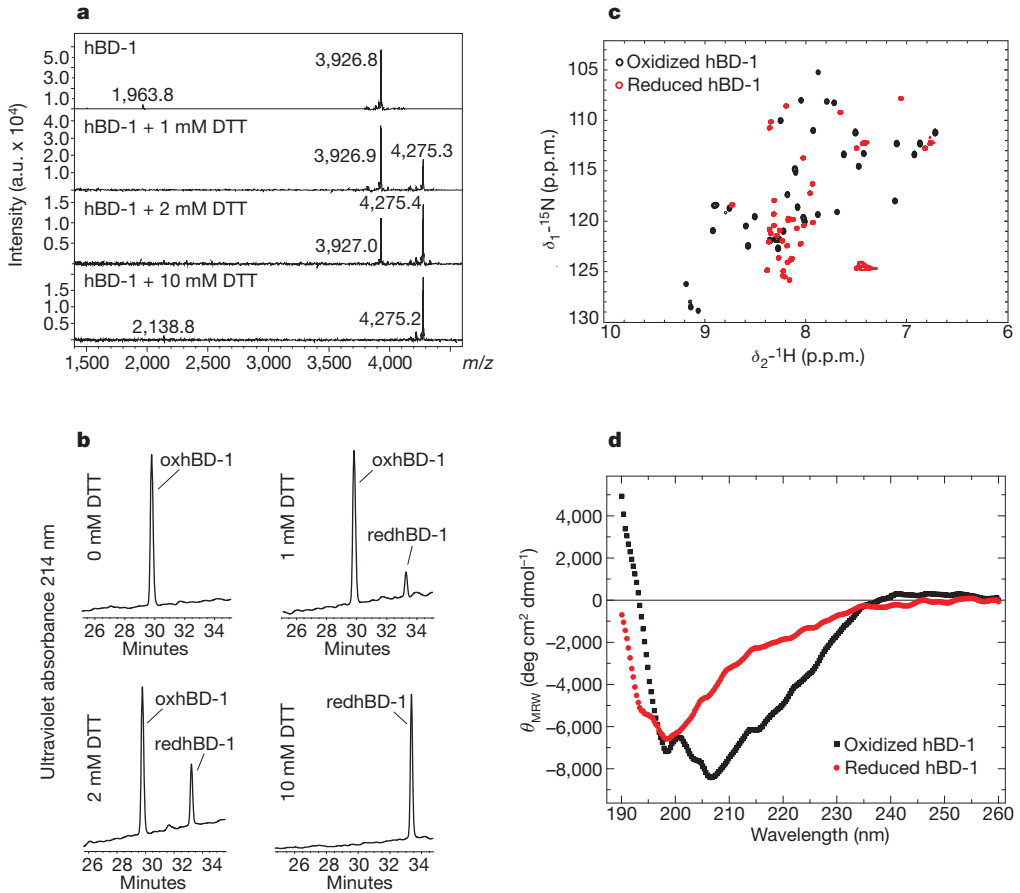


Figure 2 | Reduced hBD-1 differs structurally from oxidized hBD-1. **a**, hBD-1 (1 μg) was incubated with different concentrations of DTT, alkylated and analysed by MALDI-MS. **b**, hBD-1 (1 μg) was incubated with different concentrations of DTT and analysed by RP-HPLC. oxhBD-1, oxidized hBD-1; redhBD-1, reduced hBD-1. **c**, Nuclear magnetic resonance (NMR) analysis of hBD-1. Superimposed ^{1}H-^{15}N-heteronuclear single quantum coherence (HSQC) spectra of oxidized (black) and reduced (red) hBD-1. δ_1-^{15}N, chemical shifts of ^{15}N nuclei; δ_2-^{1}H, chemical shifts of protons. **d**, Circular dichroism spectroscopy of oxidized (black) and reduced (red) hBD-1. Θ_{MRW}, molar ellipticity normalized to mean amino acid residue weight.

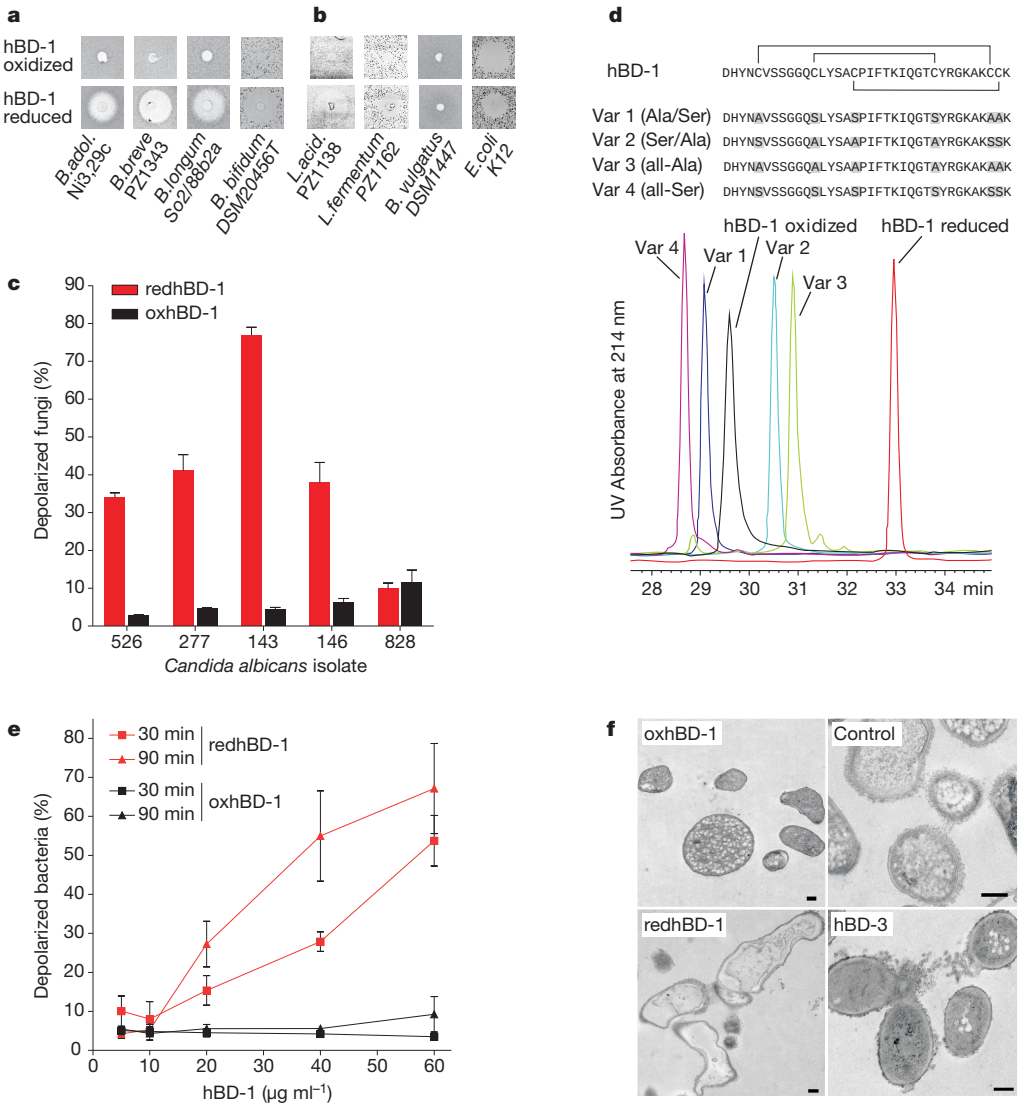


Figure 3 | Reduced but not oxidized hBD-1 has a microbicidal effect. **a**, **b**, oxhBD-1 and redhBD-1 (1.5 μg) were incubated with bifidobacteria (**a**) and lactobacilli (3 μg defensin), *Bacteroides vulgatus* and *Escherichia coli* (**b**). *B. adol.*, *B. adolescentis*; *L. acid.*, *Lactobacillus acidophilus*. **c**, Flow cytometric antimicrobial killing assay of *C. albicans* incubated with reduced (red) or oxidized (black) hBD-1. Data are presented as mean ± s.e.m. of two independent experiments each done in duplicates. **d**, RP-HPLC of oxidized and reduced hBD-1 and alanine/serine variants. **e**, Flow cytometric antimicrobial killing assay of *B. adolescentis* incubated with reduced (red) or oxidized (black) hBD-1 for 30 (squares) or 90 (triangles) minutes. Results are presented as mean ± s.e.m. of two independent experiments each done in duplicates. **f**, Transmission electron microscopy of *B. adolescentis* incubated with oxidized hBD-1, without defensin (control), reduced hBD-1 or with hBD-3. Scale bars, 200 nm.

investigate if a linear structure itself accounts for the augmented activity (Fig. 3d). All variants were either inactive or only weakly active against bifidobacteria and lactobacilli (Supplementary Fig. 4). The only bacterium being sensitive was *E. coli* K12, which was also sensitive to oxhBD-1 (Fig. 3b). Since redhBD-1 showed increased hydrophobicity, we performed RP-HPLC analyses of the Ala/Ser-variants (Fig. 3d), revealing that these eluted earlier, indicating a lower surface hydrophobicity compared to redhBD-1.

A truncated variant of hBD-1 (hBD-$1_{\Delta 30-36}$) lacking the seven C-terminal amino acids *GKAKCCK* did not inhibit growth of *B. adolescentis* under normal nor reducing conditions whereas the C-terminal seven amino acid peptide itself was antimicrobially active (data not shown). Comparable activity was observed when reversing the amino acid sequence of the C-terminal peptide while substituting cysteines by alanines or serines abolished antimicrobial activity completely. Accordingly, unfolding itself is not sufficient for antimicrobial activity of redhBD-1; rather hydrophobicity and free cysteine residues located in the C terminus are important features (Supplementary Fig. 5).

By using a flow cytometric antimicrobial killing assay[17] we found that redhBD-1 had a rapid bactericidal effect: redhBD-1, but not oxhBD-1, caused bacterial membrane depolarization of *B. adolescentis* after incubation for only 30 min (Fig. 3e). Correspondingly, when performing transmission electron microscopy with *B. adolescentis* (Fig. 3f), oxhBD-1 did not cause any morphological damage of bacterial cells. In contrast, incubation with redhBD-1 caused loss of stainable cytoplasmic content through a process which seemed to be initiated intracellularly. Hence, we assume that linear hBD-1 kills bacteria by using a mechanism being different from that of hBD-3, which primarily attacks the bacterial membrane (Fig. 3f).

Physiologically, enzymatic redox-regulation is mainly controlled by thioredoxin (TRX), a multifunctional and ubiquitously expressed oxidoreductase[11,18]. Because TRX is expressed by mucosal surfaces and shows extracellular functions[19], we tested if hBD-1 is a natural substrate for thioredoxin. Therefore we incubated oxhBD-1 with different concentrations of TRX in the presence of nicotinamide adenine dinucleotide phosphate (NADPH) and thioredoxin reductase (TrxR), composing the

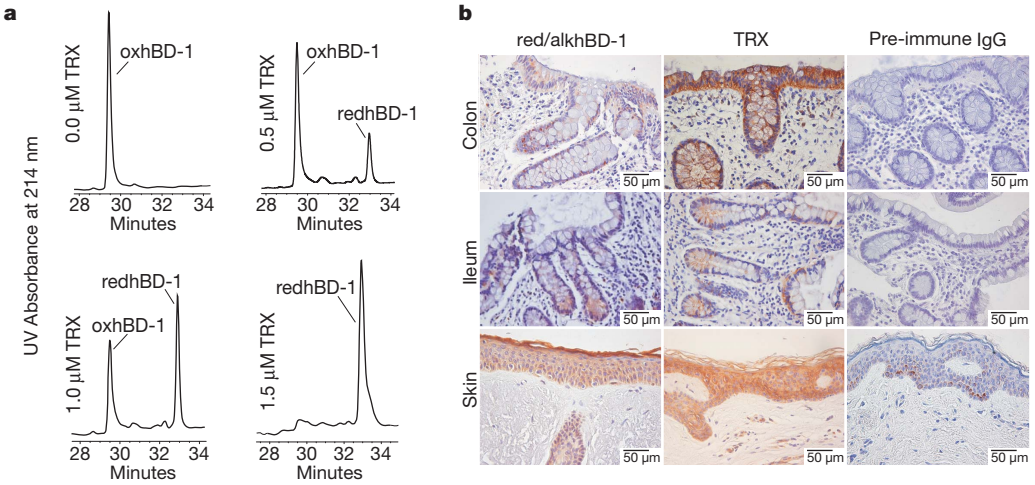


Figure 4 | Thioredoxin (TRX) catalyses reduction of oxidized hBD-1 and co-localizes with redhBD-1 *in vivo*. **a**, Oxidized hBD-1 was incubated with human thioredoxin, rat thioredoxin reductase and NADPH. Incubation mixtures were analysed by RP-HPLC. **b**, Immunohistochemical analysis of reduced/alkylated hBD-1 (red/alkhBD-1; left), thioredoxin (TRX, centre) and pre-immune IgG fraction (right). Staining of human colon (upper row), ileum (middle row) and skin sections (bottom row) are shown. In the pre-immune IgG-treated skin (bottom row, right), brownish cells in the basal epidermal layer are melanocytes. Scale bars, 50 µm.

natural thioredoxin system[11], and analysed samples by RP-HPLC to evaluate oxidized and reduced hBD-1 fractions. hBD-1 was reduced by thioredoxin in a concentration-dependent fashion (Fig. 4a). Likewise, increasing TrxR concentration increased the amount of redhBD-1 while omitting TrxR prevented reduction (data not shown). These findings were confirmed by a sensitive spectrophotometric protein disulphide reduction assay[20] measuring the consumption of NADPH (Supplementary Fig. 6). Although minor amounts of NADPH were consumed when using protein disulphide isomerase (PDI) instead of TRX or by using TrxR without TRX (data not shown), we propose that the complete thioredoxin system might be a physiological mediator catalysing effective reduction of oxhBD-1.

To document the presence of reduced hBD-1 *in vivo*, we generated a redhBD-1-specific antibody. This antibody recognized reduced hBD-1 and alkylated hBD-1, but not correctly folded synthetic or recombinant hBD-1 in an immunodot assay as well as during western blot analyses (Supplementary Fig. 7). redhBD-1 was observed in human colonic mucosa and at the bottom of small intestinal crypts by immunohistochemistry staining (Fig. 4b). Additionally, redhBD-1 staining was observed in human skin epidermis, whereas the protein G-purified pre-immune IgG fraction did not cause any staining. Analyses with a TRX-specific antibody revealed a similar staining pattern, indicating that redhBD-1 and TRX co-localize and thereby strengthened the hypothesis that TRX is a physiological mediator catalysing reduction of hBD-1 in human epithelia.

In inflamed tissue from rodent models, application of recombinant thioredoxin has a beneficial effect and ameliorates colitis[21]. We therefore studied TRX mRNA in human inflammatory bowel disease and found it to be decreased in inflamed colonic tissue (Supplementary Fig. 8). It remains unclear whether this relative lack of thioredoxin could perpetuate inflammation via a reduction of intestinal antibiotic barrier function. Unfortunately, due to differences in the antimicrobial peptide repertoire in rodents compared to humans definite proof of such a concept will be difficult to obtain.

Anaerobic niches can be found in cutaneous sweat- and sebaceous glands, wounds, infectious sites as well as mucosal membranes of intestine, vagina, oral cavities and others[22,23]. It seems plausible that hBD-1 and thioredoxin are constitutively produced by these epithelia to provide a hostile environment for (facultative) anaerobic pathogens or commensals. A disease relevance of these mechanisms appears likely, since single nucleotide polymorphisms in the hBD-1-encoding *DEFB1* promoter region have been associated with increased risk of caries, periodontitis, *Candida* infection and Crohn's Disease[24–28].

Because antimicrobial activity of hBD-1 was found to be comparably low[16,29] it was paradoxical why an organism produces high amounts of an ineffective defence molecule. Here we show that reducing disulphide bonds unmasks potent antimicrobial activity of hBD-1. Although its definite *in vivo* relevance remains to be demonstrated, these findings provide a new mechanism: redox modulation depending on enzymatic and environmental factors strongly augments antibiotic killing, a finding which opens new avenues in understanding the complex process of mucosal as well as skin host protection against commensals and pathogens.

METHODS SUMMARY

Antimicrobial assay. Antimicrobial assay was modified from ref. 12. Bacteria (except *E. coli* K12) were grown anaerobically and the assay was performed under anaerobic conditions with antimicrobial peptides or its variants/fragments and 0, 1 and 2 mM DTT.

Generation of recombinant hBD1 and its variants. Generation of hBD-1 and its variants is described in the Supplementary Methods and Supplementary Table 2.

Nuclear magnetic resonance (NMR) spectroscopy. Uniformly ^{15}N-labelled reduced and oxidized hBD-1 was solubilised in H_2O containing 10% D_2O to a final concentration of 0.5–0.7 mM. All spectra were recorded at 298 K.

Circular dichroism (CD) spectroscopy. CD spectroscopy was carried out at 25 °C in 10 mM sodium phosphate, pH 7.4. Spectra were recorded at 0.1 mg ml^{-1} in the far-ultraviolet range.

Transmission electron microscopy of bacteria. *Bifidobacterium adolescentis* was incubated with 200 µg ml^{-1} synthetic defensin. Bacteria were fixed in Karnovsky's fixative, embedded in agarose, coagulated, cut in small blocks and fixed again in Karnovsky's solution. After post-fixation and embedding in glycid ether blocks were cut using an ultra microtome. Sections (30 nm) were mounted on copper grids and analysed using a Zeiss LIBRA 120 transmission electron microscope.

Generation of anti-reduced-hBD-1-specific antibodies. Generation of anti-reduced-hBD-1-specific antibodies and its use for immunohistochemistry and dot blot is described in Methods and Supplementary Fig. 7.

Western blot analysis. Western blot analysis of reduced/alkylated hBD-1 is described in ref. 30 and Methods.

Thioredoxin reduction assay. Thioredoxin reduction assays were performed analogous to ref. 20. Briefly, oxhBD-1 was incubated with 0.5–2.0 µM thioredoxin or protein disulphide isomerase in the presence of 100 nM thioredoxin reductase and 0.8 mM NADPH in 0.1 M potassium phosphate buffer-2 mM EDTA. Incubation mixtures were analysed with RP-HPLC. For spectrophotometric disulphide reduction assay similar conditions were used in a volume of 100 µl, absorbance decrease was monitored at 340 nm.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 7 December 2009; accepted 17 November 2010.

1. Round, J. L. & Mazmanian, S. K. The gut microbiota shapes intestinal immune responses during health and disease. *Nature Rev. Immunol.* **9,** 313–323 (2009).
2. Macpherson, A. J. & Harris, N. L. Interactions between commensal intestinal bacteria and the immune system. *Nature Rev. Immunol.* **4,** 478–485 (2004).
3. Hooper, L. V. *et al.* Molecular analysis of commensal host-microbial relationships in the intestine. *Science* **291,** 881–884 (2001).
4. Wilson, M. in *Microbial Inhabitants of Humans* Ch. 7 pp. 251–317 (Cambridge University Press, 2005).
5. Zasloff, M. Antimicrobial peptides of multicellular organisms. *Nature* **415,** 389–395 (2002).
6. Harder, J., Glaser, R. & Schroder, J. M. Human antimicrobial proteins effectors of innate immunity. *J. Endotoxin Res.* **13,** 317–338 (2007).
7. Bevins, C. L. Antimicrobial peptides as effector molecules of mammalian host defense. *Contrib. Microbiol.* **10,** 106–148 (2003).
8. Peschel, A. & Sahl, H. G. The co-evolution of host cationic antimicrobial peptides and microbial resistance. *Nature Rev. Microbiol.* **4,** 529–536 (2006).
9. Bensch, K. W. *et al.* hBD-1: a novel β-defensin from human plasma. *FEBS Lett.* **368,** 331–335 (1995).
10. Tollin, M. *et al.* Antimicrobial peptides in the first line defence of human colon mucosa. *Peptides* **24,** 523–530 (2003).
11. Arnér, E. S. J. & Holmgren, A. Physiological functions of thioredoxin and thioredoxin reductase. *Eur. J. Biochem.* **267,** 6102–6109 (2000).
12. Lehrer, R. I. *et al.* Ultrasensitive assays for endogenous antimicrobial polypeptides. *J. Immunol. Methods* **137,** 167–173 (1991).
13. Ganz, T. & Lehrer, R. I. Defensins. *Curr. Opin. Immunol.* **6,** 584–589 (1994).
14. Scudiero, O. *et al.* Novel synthetic, salt-resistant analogs of human beta-defensins 1 and 3 endowed with enhanced antimicrobial activity. *Antimicrob. Agents Chemother.* **54,** 2312–2322 (2010).
15. Taylor, K., Barran, P. E. & Dorin, J. R. Structure-activity relationships in beta-defensin peptides. *Biopolymers* **90,** 1–7 (2008).
16. Nuding, S. *et al.* Antibacterial activity of human defensins on anaerobic intestinal bacterial species: a major role of HBD-3. *Microbes Infect.* **11,** 384–393 (2009).
17. Nuding, S. *et al.* A flow cytometric assay to monitor antimicrobial activity of defensins and cationic tissue extracts. *J. Microbiol. Methods* **65,** 335–345 (2006).
18. Holmgren, A. Thioredoxin. *Annu. Rev. Biochem.* **54,** 237–271 (1985).
19. Sido, B. *et al.* Potential role of thioredoxin in immune responses in intestinal lamina propria T lymphocytes. *Eur. J. Immunol.* **35,** 408–417 (2005).
20. Holmgren, A. Enzymatic reduction-oxidation of protein disulfides by thioredoxin. *Methods Enzymol.* **107,** 295–300 (1984).
21. Tamaki, H. *et al.* Human thioredoxin-1 ameliorates experimental murine colitis in association with suppressed macrophage inhibitory factor production. *Gastroenterology* **131,** 1110–1121 (2006).
22. Nizet, V. & Johnson, R. S. Interdependence of hypoxic and innate immune responses. *Nature Rev. Immunol.* **9,** 609–617 (2009).
23. Nagy, E. Anaerobic infections: update on treatment considerations. *Drugs* **70,** 841–858 (2010).
24. Ozturk, A., Famili, P. & Vieira, A. R. The antimicrobial peptide DEFB1 is associated with caries. *J. Dent. Res.* **89,** 631–636 (2010).
25. Schaefer, A. S. *et al.* A 3′ UTR transition within *DEFB1* is associated with chronic and aggressive periodontitis. *Genes Immun.* **11,** 45–54 (2010).
26. Jurevic, R. J. *et al.* Single-nucleotide polymorphisms (SNPs) in human β-defensin 1: high-throughput SNP assays and association with *Candida* carriage in type I diabetics and nondiabetic controls. *J. Clin. Microbiol.* **41,** 90–96 (2003).
27. Kocsis, A. K. *et al.* Association of beta-defensin 1 single nucleotide polymorphisms with Crohn's disease. *Scand. J. Gastroenterol.* **43,** 299–307 (2008).
28. Peyrin-Biroulet, L. *et al.* Peroxisome proliferator-activated receptor gamma activation is required for maintenance of innate antimicrobial immunity in the colon. *Proc. Natl Acad. Sci. USA* **107,** 8772–8777 (2010).
29. Singh, P. K. *et al.* Production of β-defensins by human airway epithelia. *Proc. Natl Acad. Sci. USA* **95,** 14961–14966 (1998).
30. Schröder, J. M. Purification of antimicrobial peptides from human skin. *Methods Mol. Biol.* **618,** 15–30 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Katajew, H. Löffler, C. Martensen-Kerl, C. Mehrens, J. Quitzau and A. Rose for technical assistance, B. Fehrenbacher for performing electron microscopy and H.-P. Kreichgauer, C. Schäfer, O. Müller, K. R. Herrlinger and M. Escher for collecting biopsies. Furthermore we thank M. Schwab for discussions and support, C. L. Bevins and J.-M. Schröder for discussions and reading of the manuscript and Ardeypharm GmbH for providing anaerobic bacterial strains and L. Zabel for providing *C. albicans* strains. This work was supported by Deutsche Forschungsgemeinschaft (WE 436/1-1, SCH 897/1-3 and SFB685) and the Robert-Bosch Foundation (Stuttgart, Germany). J.W. is an Emmy Noether Scholar of Deutsche Forschungsgemeinschaft.

Author Contributions B.O.S. performed antimicrobial activity assays, HPLC analyses, MALDI-MS and TRX assays, designed and evaluated experiments, generated figures and wrote the manuscript. Z.W. generated and purified recombinant hBD-1, its ^{15}N-labelled forms and hBD-1-variants, generated alkhBD-1-affinity columns and affinity-purified the red/alkhBD-1-antibody. S.N. performed flow cytometric analyses, S.G. performed NMR spectroscopy and analysed data, M.M. performed CD spectroscopy and analysed data together with J.Bu., J.Be. performed RT-PCR and M.S. was in charge of electron microscopy. E.F.S. and J.W. designed and evaluated experiments and wrote the manuscript. All authors were involved in data discussions and the final version of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.W. (jan.wehkamp@ikp-stuttgart.de).

METHODS

Bacterial and fungal strains. Bacterial strains *B. adolescentis* Ni3,29c (clinical isolate), *B. adolescentis* DSM20038T (reference strain from German Collection of Microorganisms and Cell Cultures (DSM, Germany), *B. adolescentis* PZ 4009 (clinical isolate), *Bifidobacterium breve* DSM20213T (DSM reference strain), *B. breve* PZ 1343 (from probiotic VSL#3), *B. breve* Ha6/14c (clinical isolate), *Bifidobacterium longum* DSM 20219T (clinical isolate), *B. longum* So2/88b2a (clinical isolate), *Lactobacillus acidophilus* PZ 1138 (clinical isolate), and *Lactobacillus fermentum* PZ 1162 (clinical isolate) were obtained from Ardeypharm (Germany) and *Bacteroides vulgatus* DSM1447 was obtained from the German Collection of Microorganisms and Cell Cultures (DSM). *Candida albicans* strains 143, 146 and 526 were isolated from faeces, strains 277 and 828 from tracheal secretions and provided by the Institute of Laboratory Medicine, Klinik am Eichert (Göppingen, Germany).

Antimicrobial assay. Antimicrobial radial diffusion assay for anaerobic bacteria was modified from ref. 12 (see refs 31 and 32). Bacteria were grown anaerobically (AnaeroGen, Oxoid) for 24 h at 37 °C on Columbia Blood agar plates, inoculated into liquid trypticase soy broth (TSB) medium and cultivated for another 24 h. Bacterial cultures were washed and diluted to attenuance (D_{620nm}) = 0.1, 150 µl were used for killing assay. Incubation was carried out in 10 ml of 10 mM sodium phosphate containing 0.3 mg ml^{-1} of TSB powder and 1% (w/v) low EEO-agarose (Sigma-Aldrich) with 0–2 mM DTT (Sigma-Aldrich) under anaerobic conditions with synthetic hBD-1, hBD-3 (both Peptide Institute), lysozyme (Sigma-Aldrich), hBD-1 variants (recombinantly expressed as described) or synthetic C-terminal heptapeptides (emc microcollections GmbH). Reducing milieu was monitored by addition of redox indicator resazurin (1 mg l^{-1}, Sigma-Aldrich). Reduced hBD-1 was obtained by reduction with 20 mM DTT, purification with RP-HPLC as described, drying with vacuum-centrifuge and storage under argon gas at −20 °C. An overlay-gel containing 6% (w/v) TSB powder, 1% agarose and 10 mM sodium phosphate buffer (pH 7.4) with or without DTT was poured onto the plates after 3 h and after incubation for 48 h at 37 °C (*Bifidobacterium bifidum* was incubated for up to 4 days) the diameter of inhibition zones was measured and in part statistically evaluated using GraphPad Prism 4.03 (Graphpad Software) and Student's *t*-test with $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.

Matrix-assisted laser desorption/ionization mass spectroscopy (MALDI-MS). hBD-1 (1 µg) was incubated with different concentrations of DTT in 10 mM sodium phosphate buffer, pH 7.4, for 30 min at 37 °C, followed by alkylation with 20 mM iodoacetamide for 30 min at 25 °C and co-crystallized with α-cyano-4-hydroxy cinnamic acid. MALDI-MS was carried out at an ultraflex TOF/TOF machine (Bruker).

High performance liquid chromatography (HPLC) analysis. hBD-1 (1 µg) was incubated with different concentrations of DTT in 10 mM sodium phosphate buffer, pH 7.4, for 30 min at 37 °C. HPLC analysis was carried out with an Agilent 1200 series system (Agilent) and a Vydac 218TP-C18 Column (250 × 4.6 mm, 5 µm, Grace). Gradient increased from 2% B to 35% B in 33 min (solvent A, water + 0.18% (v/v) trifluoroacetic acid (TFA); solvent B, acetonitrile + 0.15% (v/v) TFA) at 25 °C and 0.8 ml min^{-1}.

Generation of recombinant hBD1 and its variants. Generation of recombinant hBD-1 and its variants is described in the Supplementary Methods part and Supplementary Table 2.

Nuclear magnetic resonance (NMR) spectroscopy. Lyophilized uniformly ^{15}N-labelled reduced and oxidized hBD-1 was solubilised in H_2O containing 10% D_2O to a final concentration of 0.5–0.7 mM. The reduced peptide was kept under nitrogen atmosphere during measurement. All spectra, the ^{1}H-^{15}N-heteronuclear single quantum coherence (HSQC), the heteronuclear edited NNH-nuclear Overhauser enhancement spectroscopy (NOESY), the conventional ^{15}N-HSQC-NOESY and the ^{1}H-^{15}N-heteronuclear NOE spectrum, were recorded at 298 K on a Bruker AVIII-600 spectrometer (Bruker).

Circular dichroism (CD) spectroscopy. CD spectroscopy was carried out on a Jasco J-715 spectropolarimeter (Jasco Corporation, Japan). Reduced hBD-1 peptide was obtained by reduction with 20 mM DTT and purification with HPLC analysis as described. Reduced peptide was kept under argon atmosphere until CD measurement and spectra were recorded in 10 mM sodium phosphate buffer, pH 7.4 at a concentration of 0.1 mg ml^{-1} in the far-ultraviolet range at 25 °C. Spectra were deconvoluted with CDSSTR and Selcon3 algorithms.

Flow cytometric antimicrobial assay. Flow cytometric antimicrobial assay measuring membrane depolarization of bacteria and fungi was carried out as described elsewhere[17]. Briefly, 1.5×10^6 cells per ml were incubated in 1:6 diluted Schaedler broth at 37 °C with hBD-1 peptide in a final volume of 50 µl. hBD-3 served as positive control. Defensins were dissolved in 0.01% acetic acid and added to bacterial/ fungal suspensions at indicated final concentrations (40 µg ml^{-1} for *C. albicans*). Bacterial or fungal suspensions incubated with solvent (0.01% acetic acid) served as controls for viability. After 30 and 90 min the suspensions were incubated for 10 min with 1 mg ml^{-1} of the membrane potential sensitive dye [bis-(1,3-dibutylbarbituric acid) trimethine oxonol] (DiBAC4(3)) (Invitrogen, USA). Suspensions were centrifuged and the sediments were resuspended in 300 µl phosphate-buffered saline. The percentage of depolarized fluorescent bacteria or fungi in suspension was determined by a FACSCalibur flow cytometer (BD, USA) using Cell Quest software (BD) for 30 000 events per sample.

Transmission electron microscopy of bacteria. Approximately 2×10^8 CFU of *Bifidobacterium adolescentis* Ni3,29c were incubated for two hours at 37 °C in an anaerobic jar in 10 mM sodium phosphate buffer (pH 7.4) containing 0.3 mg ml^{-1} of TSB-powder in the absence or presence of 200 µg ml^{-1} synthetic oxidized hBD-3, oxidized hBD-1 (Peptide Institute Inc.) and reduced hBD-1. For transmission electron microscopy, bacteria were centrifuged, fixed in pre-warmed Karnovsky's fixative for 1h at room temperature and stored at 4 °C for 24 h. After centrifugation, the sediment was embedded in 3.5% agarose at 37 °C, coagulated at room temperature, cut in small blocks and fixed again in Karnovsky's solution for at least 1 h. Post-fixation was based on 1.0% osmium tetroxide containing 1.5% K-ferrocyanide in 0.1 M cacodylate buffer for 2 h. After embedding in glycid ether the blocks were cut using an Ultracut microtome (Reichert, Austria). Ultra-thin sections (30 nm) were mounted on copper grids and analysed using a Zeiss LIBRA 120 transmission electron microscope (Carl Zeiss, Germany) operating at 120 kV.

Generation of anti-reduced-hBD-1-specific antibodies. Polyclonal anti-red-hBD-1 antibodies were generated in rabbits against a mixture of reduced and alkylated recombinant full-length hBD-1 as antigens. For each rabbit, a total of 0.9 mg of a protein mixture including 600 µg of HPLC-purified carboxamidomethylated hBD-1 and 300 µg of HPLC-purified reduced hBD-1 was conjugated to glutaraldehyde-treated maleimide-activated keyhole limpet haemocyanin (KLH) (protein-KLH 1:1, w/w) in phosphate buffered saline (PBS, pH 7.2) and used as immunogens. Immunization was carried out four times on days 0, 14, 28 and 35. Rabbits were bled 2 weeks after the last booster. Antisera were first applied to a HiTrap Protein G HP column (GE Healthcare, Germany) to separate IgG fraction. Next, antibodies were further purified by affinity chromatography using an alkylated hBD-1-column which was prepared from 1 mg highly HPLC-purified alkylated hBD-1-pet32-fusion protein that was covalently bound to a HiTrap N-hydroxysuccinimide (NHS)-activated HP 1 ml column (GE Healthcare). To deplete any cross-reacting antibodies recognizing the correctly folded oxhBD-1, the purified antibodies were further applied to an ox-hBD-1-column, which was generated from a HiTrap NHS-activated HP 1 ml column, where correctly folded oxhBD-1 (prepared as described) was covalently bound. Specificity was tested by immunodot and western blot analyses (Supplementary Fig. 7). As antigens alkylated hBD-1, synthetic, correctly folded oxhBD-1, purified recombinant oxhBD-1 and freshly prepared redhBD-1 were used. Upon immunodot and western blot-analyses purified antibody preparations recognized specifically redhBD-1 and alkylated (alk)hBD-1, but neither synthetic, correctly folded nor purified, recombinant oxhBD-1. In addition, pre-incubation of the antibody with alkylated hBD-1 prevented immunohistochemical staining (Supplementary Fig. 7).

Thioredoxin reduction assay. Thioredoxin reduction assays were performed analogous to A. Holmgren[20]. For RP-HPLC analysis 10 µM synthetic, oxhBD-1 was incubated with 0.8 mM NADPH (biomol, Germany), 100 nM rat thioredoxin reductase (IMCO, Sweden), 0.0 – 1.5 µM human thioredoxin (Sigma-Aldrich) or up to 2.0 µM bovine protein disulphide isomerase (PDI, Sigma-Aldrich) for 30 min at 37 °C in 0.1 M potassium phosphate-2 mM EDTA, pH 7.0 buffer. Incubation mixtures were acidified with TFA, mixed with HPLC solvent and analysed with HPLC as described and conversion from oxidized into reduced hBD-1 peptide was followed by retention time. For spectrophotometric disulphide reduction assay similar conditions were used with the following exceptions: oxhBD-1 concentration was either 25 µM while TRX concentration was 0.5 – 2.0 µM or hBD-1 concentration ranged from 12.5 to 100 µM while TRX concentration was constant at 2.0 µM. Absorbance decrease was monitored at 340 nm for up to 60 min at 37 °C in a 96-well plate reader and a final volume of 100 µl. All values were corrected against a control which contained buffer instead of hBD-1; experiments were repeated at least twice.

Immunohistochemistry. Fixation of tissue samples and biopsies was performed in 4% formalin containing 20 mM iodoacetamide to conserve open disulphide bonds and a protease-inhibitor cocktail (Complete, Roche, Germany). Paraffin sections (5 µm) of tissue samples were deparaffinised and rehydrated before heat-induced antigen retrieval was performed in 0.01 M citrate buffer (pH 6.0). Slides were blocked with 12% bovine serum albumin in Tris-buffered saline, pH 7.4, before staining. Immunohistochemical staining was performed at room temperature for one hour using either affinity-purified polyclonal rabbit reduced/alkylated-hBD-1-antibody (5–20 µg ml^{-1}) or rabbit thioredoxin antibody (intestinal sections, Dako, 1:400) or goat thioredoxin antibody (skin sections, R&D Systems, 1:500). A biotinylated secondary pig anti-rabbit IgG antibody (1:300, Dako Cytomation) or rabbit anti-goat IgG antibody (1:500, Jackson Immuno Research Lab) was used,

followed by incubation with Vectastain ABC Kit Elite Pk-6100 (Vector, USA) developed with Vector NovaRED substrate kit for peroxidase Sk-4800 (Vector) and counterstained with hematoxylin. Specificity test of the reduced/alkylated hBD-1-antibody was performed by using protein G-purified pre-immune IgG (Fig. 4b) and blocking with alkylated hBD-1 (Supplementary Fig. 7).

Immunodot blot analysis. Proteins (synthetic, correctly folded oxidized hBD-1 in the absence or presence of 2 mM (tris(2-carboxyethyl) phosphine (TCEP); recombinantly expressed, refolded and HPLC-purified oxidized hBD-1 in the absence or presence of TCEP as well as HPLC-purified alkylated hBD-1) were dissolved in 0.1% (v/v) formic acid and were dotted to a Protran-nitrocellulose membrane (Schleicher & Schuell BioScience, Germany), blocked for 1 h in 5% (w/v) nonfat powdered milk in PBS + 0.05% Tween (PBST) and incubated for 18 h at 4 °C in 3% (w/v) nonfat powdered milk in PBST containing 8.8 $\mu g\,ml^{-1}$ affinity-purified polyclonal reduced/alkylated rabbit hBD-1-antibody. The membrane was washed with PBST six times for 5 min each, then incubated for 1 h in 3% (w/v) nonfat powdered milk in PBST containing 1:20.000 dilution of goat anti-rabbit IgG HRP conjugate (Dianova, Germany). After six washing steps the membrane was incubated for 5 min with chemiluminescent peroxidase substrate (Sigma) and visualized using a Diana III cooled CCD-camera imaging system (Raytest, Germany). Densitometric quantifications were performed using AIDA evaluation software (Raytest). Quality control experiments for the reduced/alkylated-hBD-1-antibody in the immunodot system (performed by pretreatment of the antibody with 5 μg/ml alkylated hBD-1) revealed no staining with oxidized, reduced as well as alkylated hBD-1 (not shown).

Western blot analysis. Samples were boiled in sampling buffer in the absence or presence of 2 mM DTT as indicated and electrophoresed in a system optimized for the analysis of 1-10 kDa-peptides as previously described for AMPs[30]. Briefly, samples were electrophoresed in a 16.5% SDS-tricine polyacrylamide gel containing 8 M urea but no reducing agent. Proteins were transferred to a Protran-nitrocellulose membrane (Schleicher & Schuell BioScience, Germany) and then treated analogue to immunodot analysis.

Isolation of total RNA and real-time PCR. Frozen biopsies were disrupted mechanically and total RNA was isolated using TRIzol reagent according to the supplier's protocol. RNA was reverse transcribed with Superscript II reverse transcriptase (Invitrogen) into cDNA, according to the standard protocol of the manufacturer. cDNA derived from 10 ng of total RNA served as a template for real-time PCR reaction. The expression levels of thioredoxin mRNA were quantified by real-time PCR using a fluorescence detection monitor (Light-Cycler; Roche Diagnostics) using oligonucleotides as shown in Supplementary Table 2. All expression levels were normalized with respect to the expression of beta-actin. To calculate statistically significant differences between groups, samples were analysed by Mann–Whitney test. Gaussian distribution was determined using the Kolmogorov–Smirnov test. Results are presented as mean ± s.e. Values of $P < 0.05$ were considered to be statistically significant. Data were analysed using Graphpad Prism version 4.

31. Schroeder, B. O. & Wehkamp, J. Measurement of antimicrobial activity under reducing conditions in a modified radial diffusion assay. *Protocol Exchange* doi:10.1038/protex.2010.204 (2011).
32. Wu, Z., Schroeder, B. O., Schroeder, J.-M. & Wehkamp, J. Production of recombinant hBD-1 in *Escherichla coli* and its specific polyclonal antibody in rabbits. *Protocol Exchange* doi:10.1038/protex.2010.205 (2011).

LETTER

doi:10.1038/nature09640

Atomic-level modelling of the HIV capsid

Owen Pornillos[1,2], Barbie K. Ganser-Pornillos[1,2] & Mark Yeager[1,2,3]

The mature capsids of human immunodeficiency virus type 1 (HIV-1) and other retroviruses are fullerene shells, composed of the viral CA protein, that enclose the viral genome and facilitate its delivery into new host cells[1]. Retroviral CA proteins contain independently folded amino (N)- and carboxy (C)-terminal domains (NTD and CTD) that are connected by a flexible linker[2–4]. The NTD forms either hexameric or pentameric rings, whereas the CTD forms symmetric homodimers that connect the rings into a hexagonal lattice[3,5–13]. We previously used a disulphide crosslinking strategy to enable isolation and crystallization of soluble HIV-1 CA hexamers[11,14]. Here we use the same approach to solve the X-ray structure of the HIV-1 CA pentamer at 2.5 Å resolution. Two mutant CA proteins with engineered disulphides at different positions (P17C/T19C and N21C/A22C) converged onto the same quaternary structure, indicating that the disulphide-crosslinked proteins recapitulate the structure of the native pentamer. Assembly of the quasi-equivalent hexamers and pentamers requires remarkably subtle rearrangements in subunit interactions, and appears to be controlled by an electrostatic switch that favours hexamers over pentamers. This study completes the gallery of substructures describing the components of the HIV-1 capsid and enables atomic-level modelling of the complete capsid. Rigid-body rotations around two assembly interfaces appear sufficient to generate the full range of continuously varying lattice curvature in the fullerene cone.

By analogy to quasi-equivalent icosahedral virus capsids, the subunits of the conical fullerene capsid of HIV-1 are organized on a hexagonal lattice that requires insertion of exactly 12 pentamers to close the shell[6,15–17]. However, the pentamers in an icosahedral capsid are in a symmetric configuration at the vertices, whereas pentamers in the fullerene cone are distributed asymmetrically, with five at the narrow end and seven at the wide end[15]. The Rous sarcoma retrovirus (RSV) CA protein can form icosahedrally symmetric particles *in vitro*, and electron cryomicroscopy maps at 10 Å resolution indicate that the hexamer and pentamer are indeed quasi-equivalent; that is, RSV CA forms both oligomers using the same interacting surfaces[9,13].

We previously used a disulphide crosslinking strategy to facilitate purification and structure determination of HIV-1 CA hexamers[11,14]. To explore the molecular basis of retroviral CA quasi-equivalence, we have now determined X-ray structures of two disulphide-crosslinked HIV-1 CA pentamers at 2.5 Å and 6 Å resolution (Supplementary Figs 1 and 2, and Supplementary Table 1). The independent pentamer structures were closely superimposable (Supplementary Fig. 3). Therefore the crosslinked proteins likely recapitulate the native pentameric architecture of HIV-1 CA. The structures were also similar to the lower-resolution structures of the RSV CA pentamer[9,13].

Comparison of the new pentamer structures with the HIV-1 CA hexamer structures[8,11,14] confirms that the two oligomers are highly related (Fig. 1). In both cases, the CTDs form a 'belt' (blue in Fig. 1a, d) surrounding the inner ring of NTDs (orange). The NTD and CTD of each CA molecule cradle the NTD from the neighbouring subunit (illustrated for the pentamer in Fig. 1b, c; for the hexamer in Fig. 1e, f; and for both oligomers in superposition in Fig. 2a). Intermolecular NTD–NTD interactions facilitate formation of the NTD rings, whereas NTD–CTD interactions hold the CTD subunits in the 'belts' against the inner NTD rings. There are no intramolecular interactions between the NTD and CTD of each subunit, apart from the covalent peptide linkage between these domains (red arrowhead in Fig. 2a). The ability of the inherently flexible linker to adopt different conformations facilitates appropriate juxtaposition of the same interaction surfaces in both the pentamer and hexamer.

Previous analysis indicated that the hexameric NTD ring closely obeys sixfold rotational symmetry and is relatively rigid, whereas the CTD subunits in the 'belts' are mobile and can rotate relative to the

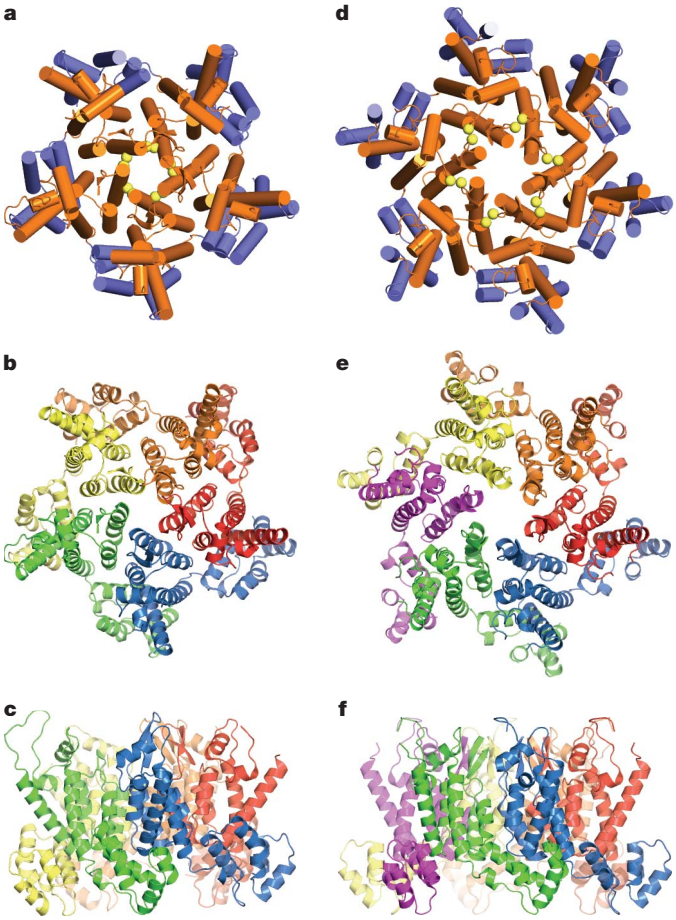


Figure 1 | Structure of the disulphide-stabilized HIV-1 CA pentamer and comparison with the hexamer. **a**, Top view of the pentamer, with the NTD coloured in orange and the CTD in blue. Helices are represented as cylinders. **b**, **c**, Top view (**b**) and side view (**c**) of the pentamer, with the helices as ribbons. Each subunit is in a different colour. **d**–**f**, Equivalent views of the hexamer (Protein Data Bank accession number 3H4E)[11]. The yellow spheres in **a** and **d** indicate the positions of the pentamer-stabilizing (N21C/A22C) and hexamer-stabilizing (A14C/E45C) disulphide bonds, respectively.

[1]Department of Molecular Physiology and Biological Physics, University of Virginia School of Medicine, Charlottesville, Virginia 22908, USA. [2]Department of Cell Biology, The Scripps Research Institute, La Jolla, California 92037, USA. [3]Division of Cardiovascular Medicine, Department of Medicine, University of Virginia Health System, Charlottesville, Virginia 22908, USA.

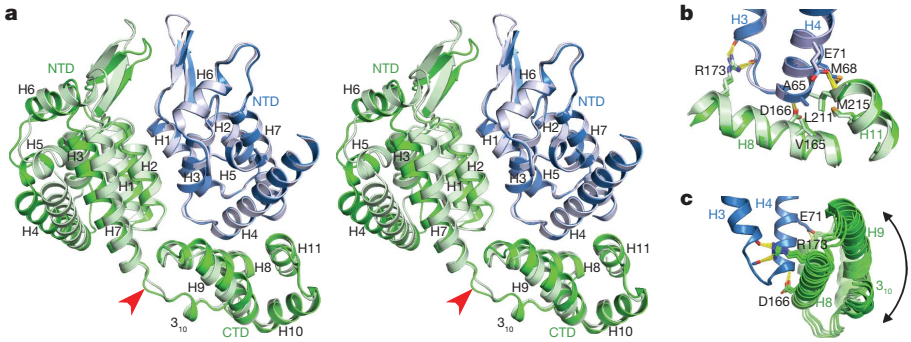


Figure 2 | Comparison of the pentamer and hexamer interactions. Both oligomers are created by quasi-equivalent packing of each of the two domains of one CA subunit (coloured in green) with the NTD of a second subunit (blue). **a**, Stereo view superposition of the pentamer interface (dark colours) and the hexamer interface (light colours). The structures are superimposed on the blue NTD. **b**, Close-up view of representative NTD–CTD contact regions in the pentamer and hexamer. Key residues are shown explicitly and labelled, with hydrogen bonds coloured in yellow. **c**, Comparison of crystallographically independent NTD–CTD interfaces in the pentamer, superimposed on the NTD. Flexion of the two domains is indicated by the black double-headed arrow, and occurs about molecular pivots composed of helix-capping intermolecular hydrogen bonds. The relevant side chains are shown explicitly and labelled, and hydrogen bonds are indicated by yellow lines.

NTD ring[11]. Each CTD pivots as a rigid body about four intermolecular helix-capping hydrogen bonds at the NTD–CTD interface. With these motions, each hexameric ring can adopt slightly different dihedral angles relative to its adjacent rings in the capsid lattice[9,11,12]. The NTD–CTD interfaces within the pentamer and hexamer are remarkably similar (Fig. 2b). Indeed, comparison of the crystallographically distinct pentamers reveals the same type of flexibility as the hexamer (that is, the fivefold symmetric NTD ring is apparently rigid, and the CTDs are more mobile) (Fig. 2c and Supplementary Fig. 3). These results indicate that the pentamer and hexamer use the same mechanism to accommodate local variations in capsid lattice curvature.

NTD ring interactions are mediated by the first three α-helices of each subunit, which form a 15-helix barrel in the pentamer (Fig. 3a) and an 18-helix barrel in the hexamer (Fig. 3b). There are subtle differences in the repeating set of NTD–NTD contacts, which comprise a three-helix bundle, with helix 2 of one subunit (orange in Fig. 3c) packed lengthwise against helices 1 and 3 of the adjacent subunit (blue). Aliphatic sidechains at the centre of the bundle form a small hydrophobic core, whereas polar residues at the periphery participate in hydrophilic interactions. As noted previously, direct hydrophilic protein–protein contacts are conspicuously absent in the hexameric NTD ring, and essentially all the intersubunit hydrogen bonds are bridged by ordered water molecules[11]. In the pentamer structure, ordered waters were not modelled (see Methods), but examination of residual difference density indicates that the assembly interfaces are likewise solvated (not shown).

The fivefold and sixfold symmetric NTD rings are distinguished through the angle subtended by adjacent subunits (72° in the pentamer and 60° in the hexamer, with the angle vertex at the centre of each ring) (Fig. 3a, b). To a first approximation, this difference is accommodated by a simple rotation of the subunits relative to each other (Fig. 3d). Remarkably, the rotation axis appears to coincide with the centre of the three-helix bundle (red dot in Fig. 3d), which allows essentially equivalent packing of the aliphatic residues. Thus the hydrophobic NTD–NTD interactions are conserved in the two rings. In contrast, the polar atoms at the outer edges of the three-helix bundle display substantially different interatomic distances, but compensatory movements of water molecules appear to maintain ring-stabilizing hydrogen bonds (not shown). Our structures therefore indicate that switching between the CA hexamer and pentamer occurs through subtle changes in intersubunit bonding interactions, and follow the principles of quasi-equivalence as originally envisioned by Caspar and Klug[18].

Pentamer formation brings charged residues at the centre of the ring in close proximity, inducing both attractive and repulsive ionic interactions. This implies that electrostatic forces control switching between the pentamer and hexamer, as was suggested for RSV CA[9]. For the HIV-1 CA oligomers, the points of closest approach occur at an annulus at the top of helix 1, which is occupied by an arginine residue (Arg 18). The arginines appear well accommodated within the annulus of the hexamer, but are more closely apposed in the pentamer (Supplementary Fig. 4). Closer juxtaposition of like charges is expected to create stronger electrostatic repulsion, which is consistent with the biochemical observation that assembly of HIV-1 CA pentamers is

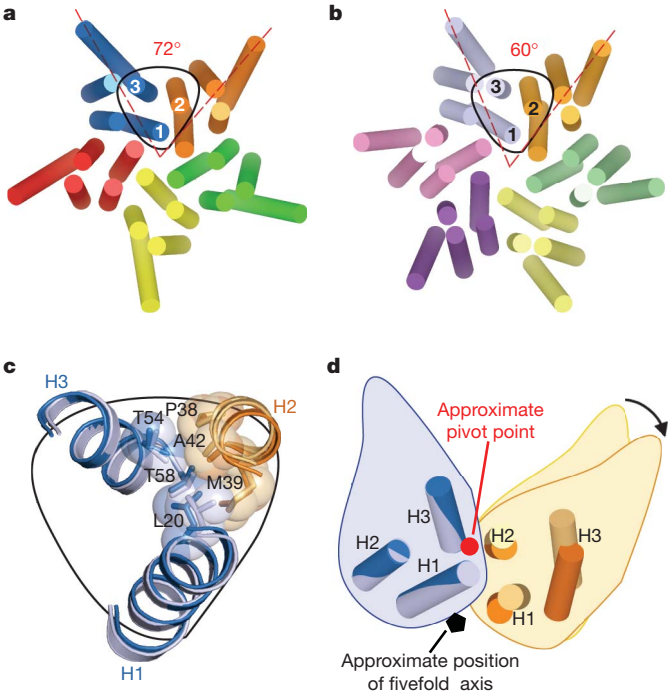


Figure 3 | Quasi-equivalence in the pentameric and hexameric NTD rings. **a**, **b**, Top views of the pentameric (**a**) and hexameric (**b**) NTD rings, with each subunit in a different colour. Subunits in the pentamer and hexamer are shown in darker and lighter shades, respectively. The angles subtended by adjacent domains are shown explicitly for the blue and orange subunits. One of the repeating three-helix units is outlined in black. **c**, Close-up view of the pentameric and hexameric repeat units, superimposed on helices 1 and 3 of the blue subunit. The aliphatic residues that form a small hydrophobic core are shown explicitly and labelled. **d**, The 'rotation' between adjacent subunits, in going from the hexamer to the pentamer. The approximate position of the rotation axis is indicated by the red dot. Note that this axis is parallel to neither the pentameric nor hexameric symmetry axes.

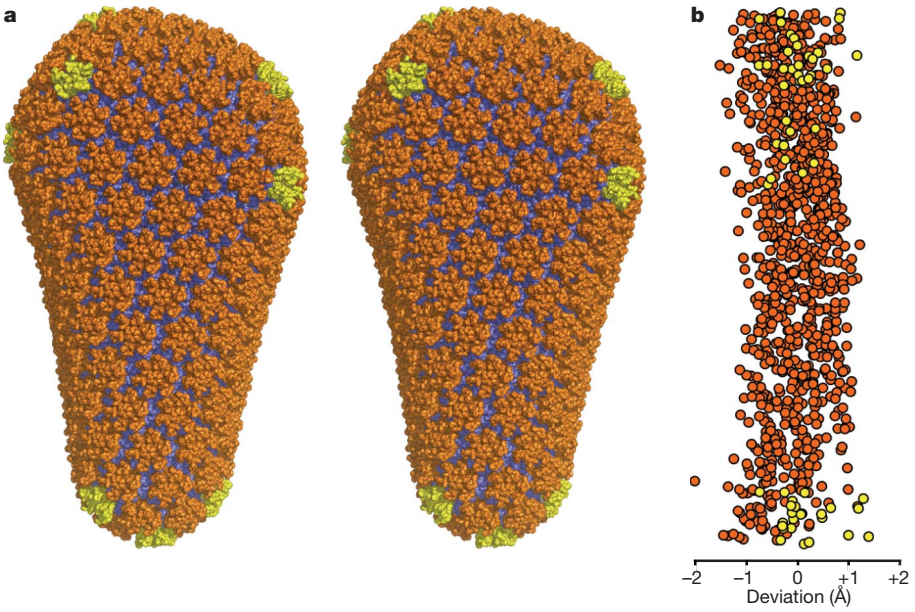


Figure 4 | Model of the HIV-1 capsid. **a**, Stereo view of a backbone-only fullerene cone model composed of 1,056 CA subunits. The hexamers, pentamers and dimers are coloured in orange, yellow and blue, respectively. Note that the capsid displays a variably curved surface. In the body of the cone, curvature changes continually, which was modelled by subunit flexion at the NTD–CTD interface. Pentamers alter the trajectory of the hexagonal lattice and create regions of sharp curvature (that is, declinations). Exactly 12 declinations are required to close a hexagonal lattice. Our modelling suggests that formation of the declinations entails a flexible CTD dimer. Note also that the CTD subunits surrounding the local threefold axes are in close proximity, consistent with the finding that this site constitutes a fourth set of capsid-stabilizing interactions[12]. **b**, The extent by which the intersubunit distances across the modelled NTD–CTD interfaces deviate from the expected value, as a function of cone length (deviation = $\text{distance}_{\text{modelled}} - \text{distance}_{\text{expected}}$, where $\text{distance}_{\text{expected}} = 9.0$ Å, which refers to the average separation of hydrogen-bonded pairs in the X-ray structures of the hexameric and pentameric NTD–CTD interfaces) (see Methods for details). Note that 99% of the NTD–CTD distances in the model are within 1 Å of the expected value. This suggests that the model is of good quality, in light of the sizeable number of constraints imposed on the subunit interactions.

disfavoured relative to hexamers. Elimination of the charge is expected to favour pentamer formation; indeed, mutation of Arg 18 into alanine promotes assembly of highly curved particles (cones, spheres, spirals, short capped cylinders)[19]. In contrast, wild-type CA typically assembles into long tubes composed of hexamers[20–22]. Mutation of Arg 18 to valine, isoleucine or leucine induces assembly of spheres (that is, more efficient pentamer formation)[8]; we presume that this is because the larger aliphatic sidechains contribute stabilizing hydrophobic contacts.

The importance of Arg 18 for the energetic landscape of HIV-1 CA assembly is consistent with its very high degree of conservation (99.8% of 2,460 sequences in the Los Alamos database, http://www.hiv.lanl.gov/content/index). We propose that electrostatic destabilization of the pentamer is precisely counterbalanced by cooperative lattice stabilization, such that pentamers form and integrate into the assembling capsid only when required to relieve strain induced by local lattice curvature. We further speculate that the narrow end of the cone may be particularly susceptible to destabilization because this region has a high concentration of pentamers, and this may be relevant to capsid disassembly or uncoating.

At this time, it is not possible to determine experimentally the atomic structure of the native HIV-1 capsid. Nevertheless, having on hand a complete gallery of high-resolution structures of the building blocks allowed us to model a fullerene cone capsid (Fig. 4a). In our modelling, the NTD hexamers, NTD pentamers and CTD dimers were treated as rigid bodies, and intersubunit distances across the NTD–CTD interfaces were used as indicators of model quality (see Methods) (Fig. 4b).

Within the body of the cone, the CA subunits were arranged on a hexagonal lattice with a unit cell spacing of approximately 93 Å. Consistent with a previous proposal[11], we found that the full range of variable lattice curvature in this region could be modelled by introducing small rigid-body rotations across the NTD–CTD interfaces, while keeping the NTD–NTD hexamerization and CTD–CTD dimerization interfaces constant. In modelling the pentameric declinations, wherein the lattice curvature is most pronounced, we found that the CTD dimers must span different distances when connecting hexamers to hexamers (approximately 33 Å) and hexamers to pentamers (approximately 26 Å) (Supplementary Fig. 5a). These distances are in close correspondence to the dimensions of two independently determined CTD dimer structures, 2KOD (ref. 12) and 1A43 (ref. 5), and suggests a rationale for the seemingly disparate structures of the dimers. Although the structures were solved from slightly different protein constructs, both retained the dimerization affinity of full-length CA[3,12]. However, each exhibited a distinct subunit packing geometry across the dimer dyad (Supplementary Fig. 5b). Therefore the 2KOD dimer was used to connect hexamers to hexamers, and the 1A43 dimer was used to connect hexamers to pentamers. Models wherein the NTD rings were connected by either dimer alone displayed significant backbone clashes (2KOD) or relatively large separations (1A43) between subunits surrounding the declinations (not shown). This suggests that rotation or slippage at the CTD–CTD interface may be a mechanistic element of capsid assembly, which is consistent with studies showing that the CTD has a flexible architecture[4,23–27].

It is remarkable that our simple modelling approach, which allowed just two types of rigid-body rotation between the building blocks, produced a fullerene model wherein essentially all the subunits displayed reasonable packing geometries. On the basis of this analysis, we conclude that CA assembly entails flexibility at both the NTD–CTD interface and the dimer interface to generate the constantly varying lattice curvature in the HIV-1 capsid.

METHODS SUMMARY

Soluble, disulphide-crosslinked pentamers of HIV-$1_{\text{NL4-3}}$ CA (containing either N21C/A22C/W184A/M185A or P17C/R18L/T19C/W184A/M185A mutations) were prepared by sequential dialysis of purified protein. Crystals were obtained by the sitting-drop vapour diffusion method in Tris-buffered precipitant solutions containing polyethylene glycol and sodium iodide. Synchrotron diffraction data were processed with the program HKL2000. Molecular replacement phasing, model building and crystallographic refinement were performed with the programs

MOLREP, Coot and PHENIX. The capsid model was built by manual rigid-body docking of the high-resolution structures of the fivefold symmetric NTD ring (Protein Data Bank accession number 3P05), the sixfold symmetric NTD ring (3H47) and twofold symmetric CTD dimers (2KOD and 1A43) into a geometric fullerene cone model.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 31 August; accepted 2 November 2010.

1. Ganser-Pornillos, B. K., Yeager, M. & Sundquist, W. I. The structural biology of HIV assembly. *Curr. Opin. Struct. Biol.* **18,** 203–217 (2008).
2. Gitti, R. K. *et al.* Structure of the amino-terminal core domain of the HIV-1 capsid protein. *Science* **273,** 231–235 (1996).
3. Gamble, T. R. *et al.* Structure of the carboxyl-terminal dimerization domain of the HIV-1 capsid protein. *Science* **278,** 849–853 (1997).
4. Berthet-Colominas, C. *et al.* Head-to-tail dimers and interdomain flexibility revealed by the crystal structure of HIV-1 capsid protein (p24) complexed with a monoclonal antibody Fab. *EMBO J.* **18,** 1124–1136 (1999).
5. Worthylake, D. K., Wang, H., Yoo, S., Sundquist, W. I. & Hill, C. P. Structures of the HIV-1 capsid protein dimerization domain at 2.6 Å resolution. *Acta Crystallogr. D* **55,** 85–92 (1999).
6. Li, S., Hill, C. P., Sundquist, W. I. & Finch, J. T. Image reconstructions of helical assemblies of the HIV-1 CA protein. *Nature* **407,** 409–413 (2000).
7. Mortuza, G. B. *et al.* High-resolution structure of a retroviral capsid hexameric amino-terminal domain. *Nature* **431,** 481–485 (2004).
8. Ganser-Pornillos, B. K., Cheng, A. & Yeager, M. Structure of full-length HIV-1 CA: a model for the mature capsid lattice. *Cell* **131,** 70–79 (2007).
9. Cardone, G., Purdy, J. G., Cheng, N., Craven, R. C. & Steven, A. C. Visualization of a missing link in retrovirus capsid assembly. *Nature* **457,** 694–698 (2009).
10. Bailey, G. D., Hyun, J. K., Mitra, A. K. & Kingston, R. L. Proton-linked dimerization of a retroviral capsid protein initiates capsid assembly. *Structure* **17,** 737–748 (2009).
11. Pornillos, O. *et al.* X-ray structures of the hexameric building block of the HIV capsid. *Cell* **137,** 1282–1292 (2009).
12. Byeon, I. J. *et al.* Structural convergence between cryoEM and NMR reveals intersubunit interactions critical for HIV-1 capsid function. *Cell* **139,** 780–790 (2009).
13. Hyun, J. K., Radjainia, M., Kingston, R. L. & Mitra, A. K. Proton-driven assembly of the Rous sarcoma virus capsid protein results in the formation of icosahedral particles. *J. Biol. Chem.* **285,** 15056–15064 (2010).
14. Pornillos, O., Ganser-Pornillos, B. K., Banumathi, S., Hua, Y. & Yeager, M. Disulfide bond stabilization of the hexameric capsomer of human immunodeficiency virus. *J. Mol. Biol.* **401,** 985–995 (2010).
15. Ganser, B. K., Li, S., Klishko, V. Y., Finch, J. T. & Sundquist, W. I. Assembly and analysis of conical models for the HIV-1 core. *Science* **283,** 80–83 (1999).
16. Jin, Z., Jin, L., Peterson, D. L. & Lawson, C. L. Model for lentivirus capsid core assembly based on crystal dimers of EIAV p26. *J. Mol. Biol.* **286,** 83–93 (1999).
17. Heymann, J. B., Butan, C., Winkler, D. C., Craven, R. C. & Steven, A. C. Irregular and semi-regular polyhedral models for Rous sarcoma virus cores. *Comput. Math. Methods Med.* **9,** 197–210 (2008).
18. Caspar, D. L. & Klug, A. Physical principles in the construction of regular viruses. *Cold Spring Harb. Symp. Quant. Biol.* **27,** 1–24 (1962).
19. Ganser-Pornillos, B. K., von Schwedler, U. K., Stray, K. M., Aiken, C. & Sundquist, W. I. Assembly properties of the human immunodeficiency virus type 1 CA protein. *J. Virol.* **78,** 2545–2552 (2004).
20. Ehrlich, L. S., Agresta, B. E. & Carter, C. A. Assembly of recombinant human immunodeficiency virus type 1 capsid protein *in vitro*. *J. Virol.* **66,** 4874–4883 (1992).
21. Campbell, S. & Vogt, V. M. Self-assembly *in vitro* of purified CA-NC proteins from Rous sarcoma virus and human immunodeficiency virus type 1. *J. Virol.* **69,** 6487–6497 (1995).
22. Gross, I., Hohenberg, H. & Kräusslich, H. G. *In vitro* assembly properties of purified bacterially expressed capsid proteins of human immunodeficiency virus. *Eur. J. Biochem.* **249,** 592–600 (1997).
23. Ternois, F., Sticht, J., Duquerroy, S., Kräusslich, H. G. & Rey, F. A. The HIV-1 capsid protein C-terminal domain in complex with a virus assembly inhibitor. *Nature Struct. Mol. Biol.* **12,** 678–682 (2005).
24. Ivanov, D. *et al.* Domain-swapped dimerization of the HIV-1 capsid C-terminal domain. *Proc. Natl Acad. Sci. USA* **104,** 4353–4358 (2007).
25. Alcaraz, L. A., del Alamo, M., Barrera, F. N., Mateu, M. G. & Neira, J. L. Flexibility in HIV-1 assembly subunits: solution structure of the monomeric C-terminal domain of the capsid protein. *Biophys. J.* **93,** 1264–1276 (2007).
26. Bartonova, V. *et al.* Residues in the HIV-1 capsid assembly inhibitor binding site are essential for maintaining the assembly-competent quaternary structure of the capsid protein. *J. Biol. Chem.* **283,** 32024–32033 (2008).
27. Wong, H. C., Shin, R. & Krishna, N. R. Solution structure of a double mutant of the carboxy-terminal dimerization domain of the HIV-1 capsid protein. *Biochemistry* **47,** 2289–2297 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This study was funded by grants from the US National Institutes of Health to M.Y. (R01-GM066087 and P50-GM082545). X-ray diffraction data were collected at beamlines 22-BM and 22-ID at the Advanced Photon Source, Argonne National Laboratory. Initial crystal screening was performed with the assistance of S. Banumathi through the Collaborative Crystallography Program, Lawrence Berkeley National Laboratory at the Advanced Light Source. We thank J. E. Johnson and D. Borek for crystallographic advice; Y. Hua for assistance with molecular biology experiments; and I. A. Wilson, C. P. Hill and W. I. Sundquist for critical reading of the manuscript.

Author Contributions All authors designed/performed the experiments, analysed the data and wrote the manuscript. O.P. performed the computational aspects of crystallographic structure determination.

Author Information The coordinates and structure factors are deposited in Protein Data Bank under accession numbers 3P05 (N21C/A22C-stabilized pentamer) and 3P0A (P17C/T19C-stabilized pentamer). To reflect the limited resolution of the second structure properly, only Cα coordinates are deposited. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.Y. (yeager@virginia.edu).

METHODS

Initial characterization of HIV-1 CA pentamer constructs. Protein expression and purification protocols have been described in detail[11,14]. The disulphide-stabilized CA pentamers were serendipitously generated in a previous screen for disulphide-stabilized hexamers[11,14]. In the hexamer study, we engineered several pairs of cysteine residues in the NTD and assayed for pairs that efficiently oxidized into intermolecular disulphide bonds linking each subunit to its two neighbours in the hexameric ring. Hexamer formation and disulphide crosslinking were induced by functional assembly of the CA mutants into helical tubes, which are composed only of hexamers[6,12]. We identified two pairs of cysteine mutations (P17C/T19C and N21C/A22C) that instead induced assembly of CA spheres approximately 35 nm in diameter (data not shown, but see Fig. 2 in ref. 14). This result implied that the assembling hexagonal lattice now incorporated pentamers[8,9,19]. Consideration of the protein mass and volume suggested that these spheres have the expected size for a $T = 3$ icosahedral particle comprising 20 hexamers and 12 pentamers, although most of the particles lacked strict icosahedral order. Non-reducing SDS–polyacrylamide gel electrophoresis of the spheres was used to confirm that the P17C/T19C and N21C/A22C mutations indeed selectively stabilized the CA pentamer (see Fig. 2 in ref. 14). As in the hexamer case[11,14], we also introduced the W184A and M185A mutations to disrupt the CTD dimerization interface[3,19,28] to prevent the crosslinked pentamers from polymerizing further (without affecting formation of the pentameric ring).

Crystallization and data collection. The mutants that crystallized contained the following mutations: N21C/A22C/W184A/M185A and P17C/R18L/T19C/W184A/M185A. The R18L mutation was added to the second construct because it promoted more efficient pentamerization (data not shown). Soluble pentamers were obtained by sequential dialysis of 10 mg ml^{-1} protein into assembly buffer (50 mM Tris pH 8, 1 M NaCl) containing 100 mM β-mercaptoethanol (βME), assembly buffer with 0.2 mM βME and, finally, 20 mM Tris pH 8, 40 mM NaCl without βME. All dialysis steps were performed at 4 °C.

Soluble pentamers were crystallized in sitting drops, by mixing protein and precipitant at a ratio of 2:1. N21C/A22C-stabilized pentamers crystallized in 0.2 M sodium iodide, 0.1 M Tris pH 8–9, 27–30% (w/v) PEG 4,000. P17C/T19C-stabilized pentamers crystallized in 0.4 M sodium iodide, 0.1 M Tris pH 7–8, 30–32% (w/v) PEG 2,000 MME. Parallelepiped-shaped crystals appeared after approximately 3 days at 25 °C and displayed maximal growth (approximately 400 μm in the longest dimension) in 2 weeks. The crystals were very osmotically sensitive. Fragments suitable for data collection were obtained by transferring the large crystals into mother liquor containing 20–30% (v/v) glycerol and applying gentle pressure on the crystals' surface with a nylon loop. Under favourable conditions, this resulted in an 'explosion' of the crystals into smaller fragments approximately 50 μm in size. The fragments were then rapidly mounted and flash-frozen in liquid nitrogen. About 1 in 50 fragments gave suitable diffraction data, which were indexed, merged and scaled with HKL2000 (ref. 29).

Structure determination of N21C/A22C-stabilized pentamers. Diffraction data to 2.5 Å resolution were collected at the Advanced Photon Source beamline 22-BM. The crystals belonged to space group $P2_1$, with five CA molecules in the asymmetric unit (Supplementary Fig. 1). The self-rotation function showed a strong fivefold non-crystallographic symmetry (NCS) axis, indicating that the CA proteins were arranged as a pentameric ring. Five NTDs and five CTDs were positioned separately with MOLREP[30]. Model building was performed with Coot[31], using NCS-averaged maps calculated separately for the NTD and CTD, and a bias-minimized map calculated with the Shake&wARP algorithm[32]. Simulated annealing refinement was performed with PHENIX[33], with the two domains defined as separate NCS groups. The test set was assigned using the thin shell method, as implemented in the program DATAMAN[34]. The model has good geometry and no residues in disallowed regions of the Ramachandran plot[35,36]. Structure statistics are summarized in Supplementary Table 1.

The asymmetric unit contained several well-ordered iodide atoms, whose positions were confirmed with an anomalous difference density map calculated from a data set collected at the Advanced Light Source beamline 5.0.2 (wavelength = 1.8 Å). Smeared anomalous density peaks were also observed at the centre of the NTD barrel, but the atomic positions were difficult to pinpoint, and these iodide densities were therefore left unmodelled (not shown). Furthermore, the iodide ions appear to have perturbed the water-mediated hydrogen-bonding network; thus we did not model water molecules. We examined putative water densities in $mF_o - DF_c$ maps calculated with phases from the final model and no NCS averaging. We found that many of the residual peaks overlapped with, or had clear correspondence to, water positions in the CA hexamer structures (not shown). We therefore believe that, like the hexamer, the pentamerization interface includes significant contributions from water-bridged hydrogen bonds.

Structure determination of P17C/T19C-stabilized pentamers. Diffraction data to 6 Å resolution were collected at the Advanced Photon Source beamline 22-ID. The crystals belonged to space group $P1$, with four pentamers in the unit cell (Supplementary Fig. 2). This structure was solved by molecular replacement with MOLREP[30], using the N21C/A22C pentamer as a search model. Owing to the limited resolution, refinement of the P17C/T19C structure in PHENIX[33] treated each NTD and CTD as separate rigid units (Supplementary Table 1).

Modelling of the mature capsid. Capsid modelling was performed by rigid-body docking of the high-resolution structures of the HIV-1 CA hexamer (3H47), pentamer (N21C/A22C structure from this study) and dimers (2KOD/NMR and 1A43/X-ray) into a geometric model of a fullerene cone (Fig. 5d, e in ref. 19). Because the geometric model was built from planar hexagons and pentagons[15,19], we first defined a corresponding plane in the hexamers and pentamers for alignment. The hexamer plane was defined as a slab, perpendicular to the sixfold symmetry axis, which contained the calculated pivot points for flexion at each of the NTD–CTD interfaces. The pivot point, in turn, was estimated by analysis of the crystallographically independent NTD–CTD interfaces in the CA hexamer structures with the program HINGEFIND[37]. Docking was then performed by least-squares alignment of marker atoms with the corners of the hexagons, which corresponded to the threefold symmetry axes of the hexagonal lattice. The pentamer was treated in the same manner. At this point, the model consisted of full-length CA hexamers and pentamers in which the sixfold and fivefold NTD–NTD and NTD–CTD interfaces were invariant and the CTD dimer interfaces were distorted. The distorted dimers were then replaced by the symmetric CTD dimer structures. This was performed by alignment of the dimers through least-squares superposition on the helix 8 Cα atoms: that is, in a manner that approximated flexion at the NTD–CTD interface. The final model comprises 166 NTD hexamers, 12 NTD pentamers, 60 1A43 CTD dimers (to connect hexamers to pentamers) and 468 2KOD dimers (to connect hexamers to hexamers).

The X-ray structures of the HIV-1 CA hexamer and pentamer revealed an invariant set of helix-capping intermolecular hydrogen bonds at the NTD–CTD interface (Arg 173:Nη – Val 59:O; Arg 173:Nη – Asn 57:O; Ala 64:N – Glu 166:Oε; and Leu 211:N – Glu 71:Oε) (Fig. 2c). These interactions were not treated explicitly during assembly of the *in silico* cone model. Therefore the extent by which the modelled distances between these pairs of residues recapitulate the experimentally observed distances can be used as a gauge of model quality. The average Cα–Cα distance for the four pairs was calculated for each NTD–CTD interface (range 7.2–10.2 Å, overall average for 1,046 interfaces 8.9 Å), and compared with the average value in the X-ray structures (range 8.7–9.4 Å, average 9.0 Å) (Fig. 4b). Deviations for 99% of the modelled NTD–CTD interfaces were within 1 Å of experimental. We judge that this level of quality assessment is sufficient at this stage of the modelling.

As noted previously[11], the NTD–CTD interface is a narrow band of contacts mediated primarily by flexible polar side chains and water-mediated hydrogen bonds. Although there is a small hydrophobic component, these interactions are mediated by flexible aliphatic side chains (for example Met 68, Met 144 and Met 215). Thus geometric fit and surface complementarity are expected to be poor indicators of model quality in this case.

28. von Schwedler, U. K., Stray, K. M., Garrus, J. E. & Sundquist, W. I. Functional surfaces of the human immunodeficiency virus type 1 capsid protein. *J. Virol.* **77,** 5439–5450 (2003).
29. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
30. Vagin, A. & Teplyakov, A. MOLREP: an automated program for molecular replacement. *J. Appl. Cryst.* **30,** 1022–1025 (1997).
31. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
32. Reddy, V. *et al.* Effective electron-density map improvement and structure validation on a Linux multi-CPU web cluster: the TB Structural Genomics Consortium Bias Removal Web Service. *Acta Crystallogr. D* **59,** 2200–2210 (2003).
33. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58,** 1948–1954 (2002).
34. Kleywegt, G. J. & Jones, T. A. xdlMAPMAN and xdlDATAMAN – programs for reformatting, analysis and manipulation of biomacromolecular electron-density maps and reflection data sets. *Acta Crystallogr. D* **52,** 826–828 (1996).
35. Vaguine, A. A., Richelle, J. & Wodak, S. J. SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr. D* **55,** 191–205 (1999).
36. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
37. Wriggers, W. & Schulten, K. Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates. *Proteins* **29,** 1–14 (1997).

LETTER

doi:10.1038/nature09647

X-ray structures of general anaesthetics bound to a pentameric ligand-gated ion channel

Hugues Nury[1,2,3,4], Catherine Van Renterghem[1,2], Yun Weng[5], Alphonso Tran[5], Marc Baaden[6], Virginie Dufresne[1,2], Jean-Pierre Changeux[2,7], James M. Sonner[5], Marc Delarue[3,4] & Pierre-Jean Corringer[1,2]

General anaesthetics have enjoyed long and widespread use but their molecular mechanism of action remains poorly understood. There is good evidence that their principal targets are pentameric ligand-gated ion channels[1,2] (pLGICs) such as inhibitory $GABA_A$ (γ-aminobutyric acid) receptors and excitatory nicotinic acetylcholine receptors, which are respectively potentiated and inhibited by general anaesthetics. The bacterial homologue from *Gloeobacter violaceus*[3] (GLIC), whose X-ray structure was recently solved[4,5], is also sensitive to clinical concentrations of general anaesthetics[6]. Here we describe the crystal structures of the complexes propofol/GLIC and desflurane/GLIC. These reveal a common general-anaesthetic binding site, which pre-exists in the apo-structure in the upper part of the transmembrane domain of each protomer. Both molecules establish van der Waals interactions with the protein; propofol binds at the entrance of the cavity whereas the smaller, more flexible, desflurane binds deeper inside. Mutations of some amino acids lining the binding site profoundly alter the ionic response of GLIC to protons, and affect its general-anaesthetic pharmacology. Molecular dynamics simulations, performed on the wild type (WT) and two GLIC mutants, highlight differences in mobility of propofol in its binding site and help to explain these effects. These data provide a novel structural framework for the design of general anaesthetics and of allosteric modulators of brain pLGICs.

Understanding the mechanism of action of general anaesthetics requires the identification of their binding site(s) within the three-dimensional structure of pLGICs. To identify such a site, we used the pH-gated bacterial homologue GLIC, a homopentameric member of the pLGIC family that was recently shown to be sensitive to general anaesthetics[6] and amenable to X-ray structure determination[4,5].

Co-crystals of GLIC were grown with propofol and apo-GLIC crystals were equilibrated in mother liquor saturated with desflurane. Diffraction data were collected up to 3.2-Å (desflurane) or 3.3-Å (propofol) resolution. In both structures, strong densities in otherwise empty Fourier $F_o - F_c$ difference maps revealed bound anaesthetics in each subunit (mean peak height $8.8 \pm 0.6\sigma$ for desflurane and $5.9 \pm 0.8\sigma$ for propofol; Supplementary Fig. 1). Both molecules were found to bind in the same region, with little change in the protein conformation compared with apo-GLIC[4], a feature observed also for some soluble proteins complexed with general anaesthetics[7–9]. The binding site is located in the upper half of the transmembrane domain (Fig. 1a) in a cavity that exists within each subunit of the apo-structure. The general-anaesthetic cavity is accessible from the lipid bilayer and progressively narrows down towards the interior of the subunit (Fig. 1b, c). Another cavity of comparable volume is located at the interface between subunits, on the other side of M1. It is not accessible from the outside. A narrow tunnel (less than 3 Å in diameter) links both cavities.

Residues from a single subunit border the general-anaesthetic cavity (Fig. 2), with contributions from M1 (I201, I202, M205 and L206), M2 (the back wall, V242), M3 (Y254, T255, I258 and I259), M4 (N307 and F303, near the mouth), and from the β6–β7 loop (Y119, P120, F121, constituting the roof). Desflurane is buried deep inside the cavity and is engaged in mainly hydrophobic interactions with M1 (I201, I202), M3 (T255 and I258) and M2 (V242). Its oxygen atom is within hydrogen-bond distance of the T255 hydroxyl group. Significant additional electron density is observed in the protein neighbourhood and is attributed to lipids, with an alkyl chain obstructing the cavity entrance (Fig. 1b), as observed with the apo structure. This defines a cavity volume of 238 $Å^3$ whereas the volume of desflurane is 94 $Å^3$. Propofol lies closer to the entrance of the general-anaesthetic cavity and would clash with the lipid seen in the apo and desflurane structures. Accordingly, the presence of propofol is associated with local acyl chain reorganization. Propofol is sandwiched between M1 and M3 and interacts mainly with T255 and Y254, by van der Waals contacts (Fig. 2). In the orientations that best fit the density maps, the propofol hydroxyl group could form a hydrogen bond with Y254. Propofol lies 6 Å away from the V242 side chain, whereas desflurane is 3.5 Å away.

We have recently shown that GLIC activation is inhibited by most general anaesthetics at clinical concentrations[6]. To check whether the general-anaesthetic binding sites contribute to this inhibition, we mutated key general anaesthetics-binding residues into alanine, or into more bulky residues (mutants I202A,W,Y, V242M,W and T255A), and studied the functional effect by two-electrode voltage-clamp electrophysiology in oocytes.

Among the mutants tested, I202Y and T255A produce a marked gain of function, with a tenfold shift of the proton dose–response curve towards lower concentrations ($pH_{50} = 6.1 \pm 0.1$ with Hill number (n_H) $= 2.0 \pm 0.2$ and $pH_{50} = 6.0 \pm 0.2$ with $n_H = 1.1 \pm 0.2$ respectively), compared with WT ($pH_{50} = 5.0 \pm 0.3$ with $n_H = 1.8 \pm 0.3$) (Fig. 3a, b). T255A shows also slower apparent rate constants for activation and deactivation. The other mutants are activated by protons in a manner similar to WT, except that V242W yielded no current (Supplementary Table 2).

The inhibitory action of general anaesthetics was measured around the one-fifth maximum effective concentration (EC_{20}) of proton activation. On WT, propofol and desflurane produced 100% maximal inhibition with half-maximum inhibitory concentration (IC_{50}) values of $24 \pm 6.3\,\mu M$ ($n_H = 1.1 \pm 0.2$) and $27 \pm 13\,\mu M$ ($n_H = 0.3 \pm 0.2$) respectively. Screening the mutants for inhibition by 10 μM propofol and 500 μM desflurane shows no change when position 202 is mutated, but shows that V242M and T255A produce a parallel tenfold shift of the propofol dose–inhibition curve to lower concentration (Fig. 3c, d and Supplementary Table 3). In contrast, V242M has no effect on desflurane inhibition, whereas T255A produces a tenfold shift of the desflurane inhibition curve towards higher concentrations. Measurement of general-anaesthetic inhibition at different pH shows that both anaesthetics are more efficient at higher pH, for both WT and T255A (Fig. 3e and Supplementary Fig. 2). Strikingly, T255A increases the

[1]Institut Pasteur, Groupe Récepteurs-Canaux, F-75015 Paris, France. [2]CNRS, URA2182, F-75015 Paris, France. [3]Institut Pasteur, Unité de Dynamique Structurale des Macromolécules, F-75015 Paris, France. [4]CNRS, URA2185, F-75015 Paris, France. [5]Department of Anesthesia and Perioperative Care, University of California, San Francisco 94143, USA. [6]Institut de Biologie Physico-Chimique, CNRS UPR 9080, F-75005 Paris, France. [7]Collège de France, F-75005 Paris, France.

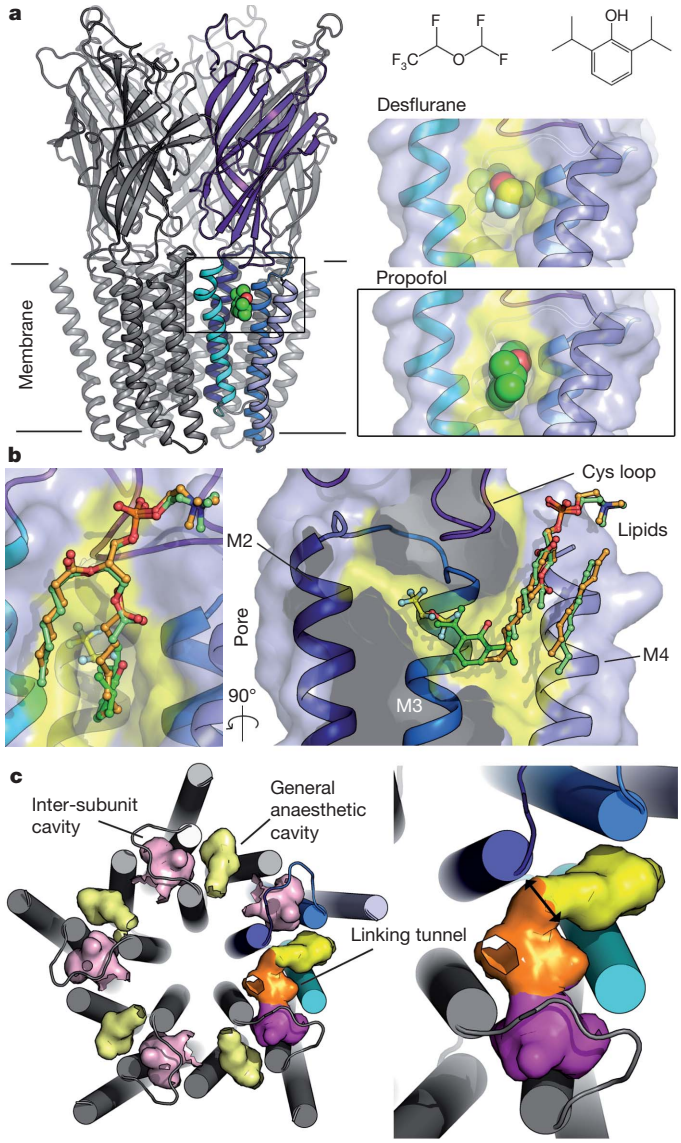


Figure 1 | Propofol and desflurane binding sites. **a**, General view of GLIC from the plane of the membrane in cartoon representation with a bound general-anaesthetic molecule in space-filling representation. The molecular surface is represented in the insets and coloured in yellow for the binding pocket. **b**, Cartoon and surface representation of the general-anaesthetic cavity seen from the membrane (left) and from the adjacent subunit (right, M1 removed for clarity) with propofol (green), desflurane (yellow) and lipids of the two structures (green and orange respectively) depicted as sticks. For this representation Cα atoms were superimposed with a root mean square deviation of 0.13 Å. **c**, Molecular surface of the general-anaesthetic intra-subunit cavities (yellow) and neighbouring inter-subunit cavities (pink) for the whole pentamer. In one of the subunits, the communication tunnel between the two cavities is depicted in orange, and its constriction indicated by an arrow in the inset.

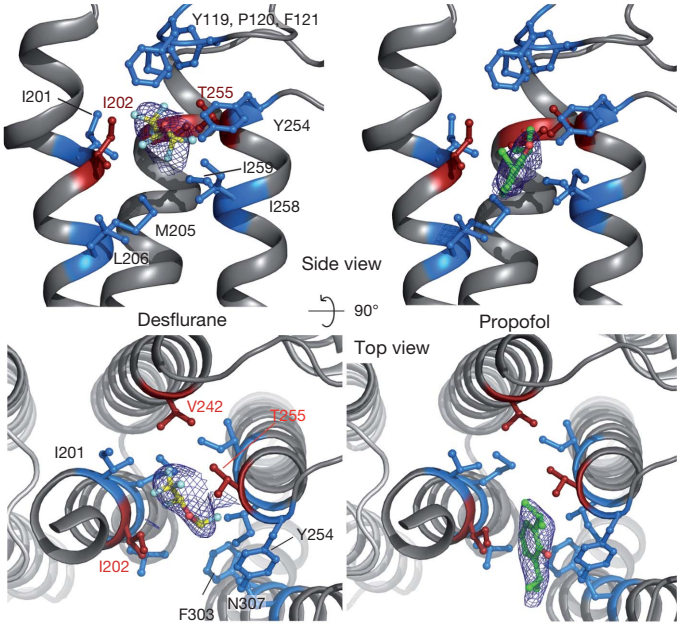


Figure 2 | Residues of the binding site. Sites for propofol (right) and desflurane (left), viewed from the membrane (top panels with M4 helix removed), and from the ECD domain (lower panels with ECD removed). Residues bordering the pocket and contributing to binding are depicted as blue or red (mutated positions) sticks. SigmaA weighted Fourier difference maps $2F_o - F_c$ contoured at 1.5σ around the anaesthetics molecules are represented as a blue mesh.

inhibition by propofol but decreases the inhibition by desflurane at all proton concentrations. Altogether, mutation of selected residues within the general-anaesthetic binding site affects (1) the intrinsic ionic response of GLIC, illustrated by the marked gain of function of I202Y and T255A, whose phenotypes are similar to that of the canonical I233(9′)A mutation[3,10], and (2) the pharmacology of general anaesthetics, illustrated both by V242M, which displays an increased sensitivity to propofol but not to desflurane, and T255A, which has an increased sensitivity to propofol but a decreased sensitivity to desflurane.

These data support the hypothesis that the general-anaesthetic binding site described here contributes to general-anaesthetic-mediated inhibition of GLIC. For desflurane, mutagenesis data match the characteristics of its binding site in the X-ray structure well, with no significant effect when the relatively distant positions 202 and 242 are mutated, and a strong impairing effect when mutating T255, which extensively contacts desflurane, into an alanine. In contrast, for propofol, both positions 202 and 255 contact propofol, but only mutation at position 255 alters its effect. More surprisingly, position 242 is not in direct contact with propofol but V242M modifies its response. These data suggest a significant mobility of propofol within the cavity, a feature that may be reflected by the high B factors of general anaesthetics in the crystal structure ($B_{desflurane} = 121$ Å^2, $B_{propofol} = 135$ Å^2, mean values), although high B factors and partial occupancy of the site cannot be discriminated at 3.3-Å resolution.

To examine this possibility further, we performed 30-ns molecular dynamics simulations of propofol bound to the WT protein, T255A, V242M and I202A mutants. At this timescale, propofol remains in the cavity, but shows substantial mobility (Fig. 4a). T255A and V242M are associated with (1) reduced propofol fluctuation (root mean square fluctuation of propofol non-H atoms of 3 ± 1.1 Å, 2.4 ± 0.8 Å, 2.3 ± 0.8 Å, 2.7 Å for the WT, T255A, V242M and I202A runs, respectively), (2) deeper penetration inside the cavity (Fig. 4b) and (3) more frequent interaction with residue 242 compared with the WT and I202A (data not shown). Altogether, these simulations provide complementary interpretations to account for the higher sensitivity of T255A and V242M to propofol inhibition that could not have been deduced from the static structure alone.

The X-ray structure of GLIC was formerly interpreted in terms of an apparently open conformation[4,5]. But general anaesthetics behave as inhibitors of the ionic response and are therefore expected to stabilize a closed conformation. Our data unravel a general-anaesthetic site in the open conformation, and molecular dynamics simulations show that propofol and desflurane are stable in this site conformation at the 30-ns timescale. This apparent contradiction can be readily explained by a non-exclusive (differential) binding of general anaesthetics to the open and closed states, with general anaesthetics displaying a higher affinity for the closed state than for the open one[11]. Interestingly, the T255A and I202Y gain-of-function phenotypes suggest a structural rearrangement

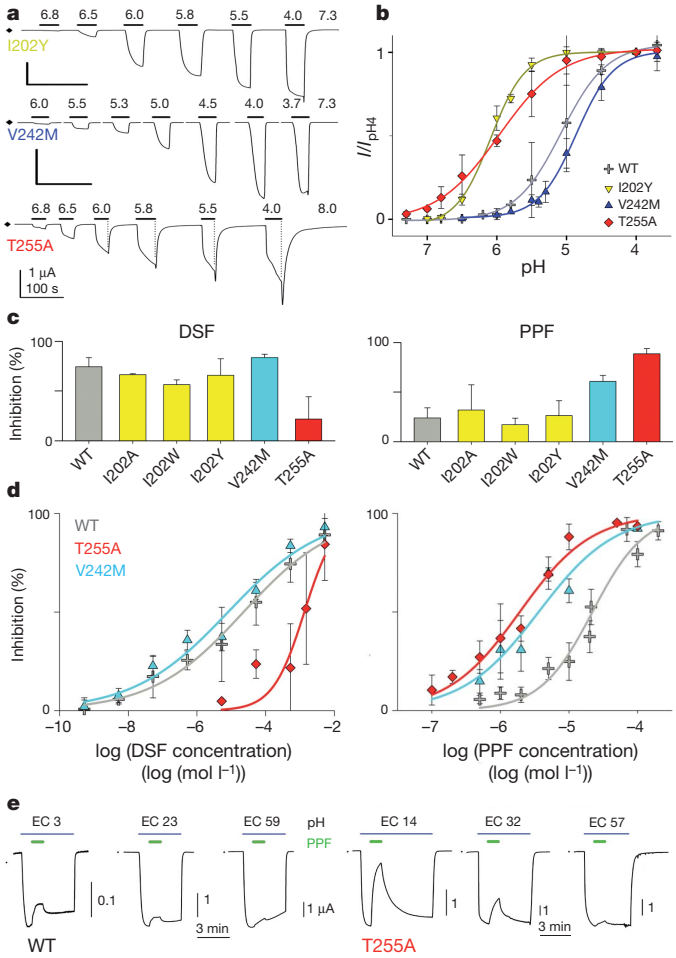


Figure 3 | Electrophysiological characterization of binding-site residues. **a**, Traces of currents evoked by 30-s applications of low extracellular pH separated by 30–60 s wash. **b**, Corresponding plots for currents normalized with respect to the value at pH 4. Mean ± s.d. of 4 to 12 cells per construct. **c**, Inhibition by 0.5 mM desflurane (left) or 10 μM propofol (right) applied for 60 s during the plateau of GLIC activation by a pH near EC_{20} ($EC_{10–30}$). **d**, Corresponding concentration–inhibition characteristics of desflurane (left) or propofol (right). **e**, Current traces showing the effect of 10 μM propofol on GLIC currents corresponding to proton $EC_{3,23,59}$ (WT, left traces) or $EC_{14,32,57}$ (T255A, right traces) of each cell. *Xenopus laevis* oocytes, holding potential −60 to −40 mV.

of the general-anaesthetic binding site during gating, in line with its location on the backside of the pore-lining M2 helices, at the level of the gate, and close to the transmembrane-domain–extracellular-domain interface. We note that the hypothetical gating mechanism[4] previously suggested from the comparison of GLIC and ELIC[12] structures involves a strong reorganization of the general-anaesthetic binding site as a consequence of M2 and M3 helix tilting.

Another important feature of the GLIC–general anaesthetic structures is that binding occurs to a site where nearby ordered lipids are identified. One lipid lying in the crevice between M1 and M4 is observed in the apo and desflurane structures and is displaced in the propofol structure. These lipids co-purify with the solubilized protein, indicating tight binding within protein subsites known to be critical for the transmembrane domain structure[13,14]. It is known that lipids contribute to pLGIC function[15,16], and those observed in the present electron-density maps are good candidates for such a role. A perturbation of interactions with lipids caused by general-anaesthetic binding might thus contribute to the functional inhibition, suggesting that general anaesthetics may compete with endogenous allosteric modulators[17], lipids in this case, but also possibly fatty acids, cholesterol and/or neurosteroids[18] in the case of eukaryotic pLGICs.

It is striking that the pharmacology of GLIC inhibition resembles that of nicotinic acetylcholine receptors (nAChRs), which are also inhibited by general anaesthetics and are unusually sensitive to volatile anaesthetics[19]. The general-anaesthetic binding site described here is thus a primary candidate for promoting nAChR inhibition. In contrast, Gly/$GABA_A$ receptors are mostly potentiated by general anaesthetics. Experimental data involving chimaeric constructs show that GlyRα1 S267 (M2), A288 (M3) and I229 (M1) contribute to general anaesthetic and alcohols potentiation[2,20]. The general anaesthetic etomidate labelled brain $GABA_A$ receptors at residues α1M236 and α3M286[21], the latter corresponding to GlyRα1 A288. This labelling was only partly inhibited by propofol and neurosteroids[22], consistent with an allosteric interaction between several binding sites. Overall, the interpretation of these data using homology models based on the cryoelectron-microscopy nAChR structure at medium resolution[23–25] suggests that intra-subunit and/or inter-subunit sites in the upper part of the transmembrane domain mediate general-anaesthetic modulation of anionic pLGICs.

The sequence of GLIC can be readily aligned with that of $GABA_AR$ and GlyR (Supplementary Fig. 3). From the GLIC three-dimensional structure and this alignment, the three residues identified in $GABA_A$/Gly receptors can be seen to point towards the inter-subunit cavity (Fig. 1c) close to the intra-subunit general-anaesthetic binding site described here. A marked reorganization of these cavities was observed in the initial steps of channel closing during our 1-μs molecular dynamics simulation of GLIC at neutral pH[26] (Supplementary Fig. 4). This involves transient communications between the inter- and intra-subunit cavities caused by M2 and M3 motions, which further supports the notion that the shape and volume of the cavity are coupled to channel gating, and could suggest that a cross-talk between both cavities might underlie general-anaesthetic-mediated potentiation.

pLGIC and particularly nAChRs are not only the target of general anaesthetics[27] but also of natural and synthetic allosteric modulators that are developed for their therapeutic potential[28,29]. Mutational data suggest that ivermectine[30] and PNU-120596, which behave as positive

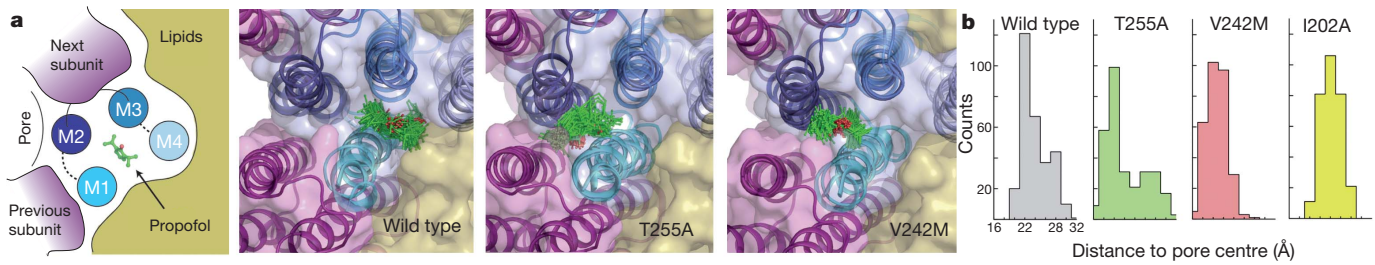


Figure 4 | Molecular dynamics simulation of propofol bound to GLIC. **a**, View from the ECD domain depicting propofol positions as green sticks, superposed at 0.5-ns intervals onto the starting (cartoon and molecular surface) and final (transparent cartoon) conformation of the protein, during the 30-ns molecular dynamics runs. The scheme on the left explains the colour code, and the three panels correspond to the WT, T255A and V242M systems. **b**, Distance distribution between the propofol centre of mass and the pore centre. The distribution is shifted closer to the pore centre for the mutants showing an increased inhibition.

allosteric modulators of α7 nAChR, bind at a location resembling that observed in our structures. The general-anaesthetic binding site unravelled herein thus provides a novel template for the design of allosteric modulators inhibiting or potentiating pLGICs.

METHODS SUMMARY

GLIC production[4], electrophysiology[3] and molecular dynamics[26] were performed as described (full methods in the Supplementary Information). Crystals were typically grown in 12–16% PEG 4000, 400 mM NaSCN, 100 mM Na-Acetate at pH 4, in the presence of an excess of general anaesthetics.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 29 April; accepted 1 November 2010.

1. Franks, N. P. General anaesthesia: from molecular targets to neuronal pathways of sleep and arousal. *Nature Rev. Neurosci.* **9,** 370–386 (2008).
2. Lobo, I. A. & Harris, R. A. Sites of alcohol and volatile anaesthetic action on glycine receptors. *Int. Rev. Neurobiol.* **65,** 53–87 (2005).
3. Bocquet, N. *et al.* A prokaryotic proton-gated ion channel from the nicotinic acetylcholine receptor family. *Nature* **445,** 116–119 (2007).
4. Bocquet, N. *et al.* X-ray structure of a pentameric ligand-gated ion channel in an apparently open conformation. *Nature* **457,** 111–114 (2009).
5. Hilf, R. J. C. & Dutzler, R. Structure of a potentially open state of a proton-activated pentameric ligand-gated ion channel. *Nature* **457,** 115–118 (2009).
6. Weng, Y., Yang, L., Corringer, P. J. & Sonner, J. M. Anaesthetic sensitivity of the *Gloeobacter violaceus* proton-gated ion channel. *Anesth. Analg.* **110,** 59–63 (2010).
7. Bhattacharya, A. A., Curry, S. & Franks, N. P. Binding of the general anaesthetics propofol and halothane to human serum albumin. High resolution crystal structures. *J. Biol. Chem.* **275,** 38731–38738 (2000).
8. Zhang, H., Astrof, N. S., Liu, J., Wang, J. & Shimaoka, M. Crystal structure of isoflurane bound to integrin LFA-1 supports a unified mechanism of volatile anaesthetic action in the immune and central nervous systems. *FASEB J.* **23,** 2735–2740 (2009).
9. Vedula, L. S. *et al.* A unitary anaesthetic binding site at high resolution. *J. Biol. Chem.* **284,** 24176–24184 (2009).
10. Revah, F. *et al.* Mutations in the channel domain alter desensitization of a neuronal nicotinic receptor. *Nature* **353,** 846–849 (1991).
11. Rubin, M. M. & Changeux, J. P. On the nature of allosteric transitions: implication of non-exclusive ligand binding. *J. Mol. Biol.* **21,** 265–274 (1966).
12. Hilf, R. J. C. & Dutzler, R. X-ray structure of a prokaryotic pentameric ligand-gated ion channel. *Nature* **452,** 375–379 (2008).
13. Haeger, S. *et al.* An intramembrane aromatic network determines pentameric assembly of Cys-loop receptors. *Nature Struct. Mol. Biol.* **17,** 90–98 (2010).
14. Villmann, C. *et al.* Functional complementation of *Glra1*$^{spd\text{-}ot}$, a glycine receptor subunit mutant, by independently expressed C-terminal domains. *J. Neurosci.* **29,** 2440–2452 (2009).
15. Nievas, G. A. F., Barrantes, F. J. & Antollini, S. S. Modulation of nicotinic acetylcholine receptor conformational state by free fatty acids and steroids. *J. Biol. Chem.* **283,** 21478–21486 (2008).
16. da Costa, C. J. B. *et al.* Anionic lipids allosterically modulate multiple nicotinic acetylcholine receptor conformational equilibria. *J. Biol. Chem.* **284,** 33841–33849 (2009).
17. Franks, N. P. & Lieb, W. R. Do general anaesthetics act by competitive binding to specific receptors? *Nature* **310,** 599–601 (1984).
18. Hosie, A. M., Wilkins, M. E., da Silva, H. M. A. & Smart, T. G. Endogenous neurosteroids regulate $GABA_A$ receptors through two discrete transmembrane sites. *Nature* **444,** 486–489 (2006).
19. Violet, J. M., Downie, D. L., Nakisa, R. C., Lieb, W. R. & Franks, N. P. Differential sensitivites of mammalian neuronal and muscle nicotinic acetylcholine receptors to general anesthetics. *Anesthesiology* **86,** 866–874 (1997).
20. Mihic, S. J. *et al.* Sites of alcohol and volatile anaesthetic action on $GABA_A$ and glycine receptors. *Nature* **389,** 385–389 (1997).
21. Li, G. *et al.* Identification of a $GABA_A$ receptor anaesthetic binding site at subunit interfaces by photolabeling with an etomidate analog. *J. Neurosci.* **26,** 11599–11605 (2006).
22. Li, G., Chiara, D. C., Cohen, J. B. & Olsen, R. W. Numerous classes of general anaesthetics inhibit etomidate binding to γ-aminobutyric acid type A ($GABA_A$) receptors. *J. Biol. Chem.* **285,** 8615–8620 (2010).
23. Miyazawa, A., Fujiyoshi, Y. & Unwin, N. Structure and gating mechanism of the acetylcholine receptor pore. *Nature* **423,** 949–955 (2003).
24. Trudell, J. R. & Bertaccini, E. Comparative modeling of a $GABA_A$ α1 receptor using three crystal structures as templates. *J. Mol. Graph. Model.* **23,** 39–49 (2004).
25. Bali, M., Jansen, M. & Akabas, M. H. GABA-induced intersubunit conformational movement in the $GABA_A$ receptor α1M1–β2M3 transmembrane subunit interface: experimental basis for homology modeling of an intravenous anaesthetic binding site. *J. Neurosci.* **29,** 3083–3092 (2009).
26. Nury, H. *et al.* One-microsecond molecular dynamics simulation of channel gating in a nicotinic receptor homologue. *Proc. Natl Acad. Sci. USA* **107,** 6275–6280 (2010).
27. Ziebell, M. R., Nirthanan, S., Husain, S. S., Miller, K. W. & Cohen, J. B. Identification of binding sites in the nicotinic acetylcholine receptor for [3H]azietomidate, a photoactivatable general anaesthetic. *J. Biol. Chem.* **279,** 17640–17649 (2004).
28. Taly, A., Corringer, P. J., Guedin, D., Lestage, P. & Changeux, J. P. Nicotinic receptors: allosteric transitions and therapeutic targets in the nervous system. *Nature Rev. Drug Discov.* **8,** 733–750 (2009).
29. Bertrand, D. & Gopalakrishnan, M. Allosteric modulation of nicotinic acetylcholine receptors. *Biochem. Pharmacol.* **74,** 1155–1163 (2007).
30. Krause, R. M. *et al.* Ivermectin: a positive allosteric effector of the α7 neuronal nicotinic acetylcholine receptor. *Mol. Pharmacol.* **53,** 283–294 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the Commission of the European Communities (Neurocypres project; to H.N.), the Louis D. Foundation of the Institut de France, the Network of European Neuroscience Institutes (ENI-NET) and a National Institutes of Health grant NIGMS R01 GM069379 (to J.M.S.). We thank J. Brallet for the gift of desflurane, the European Synchrotron Radiation Facility and Soleil staff for assistance during data collection, and G. Brannigan for providing propofol simulation parameters. Simulations were performed using high-performance computing resources from the Grand Equipement National de Calcul Intensif, Institut du Développement et des Ressources en Informatique Scientifique (GENCI-IDRIS, grant 2009-072292).

Author Contributions All authors contributed extensively to the work presented in this paper.

Author Information The coordinates of models are deposited in Protein Data Bank under accession numbers 3P50 (propofol) and 3P4W (desflurane). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to P.-J.C. (pjcorrin@pasteur.fr) or M.D. (delarue@pasteur.fr).

METHODS

Protein production. The protein was produced and purified as described previously[3,4,26] with a few variations. The GLIC protein was overexpressed in *Escherichia coli* C43 cells, and expression was induced with 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) at absorbance ($A_{600\ nm}$) = 1 overnight at 20 °C. Cells were mechanically lysed in buffer 1 (Tris 20 mM pH 7.6, NaCl 300 mM, with proteases inhibitors from Roche); membranes were isolated by ultracentrifugation. The proteins were extracted from membranes with 2% DDM (Anatrace) under agitation at 4 °C, and the solubilized fraction was cleared by ultracentrifugation. Solubilized proteins were first purified by affinity chromatography on an amylose resin. After extensive wash, in buffer 1 supplemented with 0.1% DDM, the MBP–GLIC fusion protein bound to the resin was cleaved overnight at 4 °C under gentle agitation, in the presence of three units of thrombin (Calbiochem) per 50 μg of protein and of 2 mM of $CaCl_2$. The digested protein was eluted in buffer 1 supplemented in 0.02% DDM and concentrated. It was then subjected to size exclusion chromatography on a Superose 6 10/300 GL column (GE Healthcare) equilibrated in the same buffer. Fractions of the peak corresponding to the pentamer were pooled and concentrated for crystallization experiments.

Crystallography. GLIC was crystallized using the vapour diffusion method in hanging drops at 20 °C. The concentrated (8–12 mg ml^{-1}) protein was mixed in a 1:1 ratio with reservoir solution containing typically 12–16% PEG 4000, 400 mM NaSCN and 0.1 M NaAc at pH 4.0. Crystals of the protein grew overnight. A saturating amount of desflurane was then added in the well (typically the well of a Linbro plate was completely filled). After 2–5 days of equilibration the crystals were rapidly transferred in the mother solution supplemented with 20% glycerol for cryoprotection and flash-frozen in liquid nitrogen. Pure propofol was added at the time of crystallization right after the mixture of the protein solution with the mother liquor. An emulsion was formed and small crystals grew in 2 days, nearby the propofol droplets included in the crystallization drops. Crystals were flash-frozen in the usual manner.

A great number of data sets of frozen single crystals were collected on beamlines Proxima-I of the Soleil Synchrotron and ID14 of the European Synchrotron Radiation Facility and processed with XDS[31] and CCP4 (ref. 32) programs. Crystals were isomorphous to the WT ones, with a C2 space group and one pentamer in the asymmetric unit. Non-crystallographic symmetry (NCS) restraints were used throughout the refinement performed by alternate cycles of manual building in COOT[33] and automatic refinement using REFMAC[34] and BUSTER[35]. Difference Fourier $F_o - F_c$ maps were checked for strong signals indicating the presence of a ligand. In the two data sets presented here, bound anaesthetics corresponded to 5–9σ peaks (Supplementary Fig. 1), depending on subunits. The peaks were present in each subunit in a region devoid of any density in other data sets. Moreover, the electron density for the known bound detergents in the pore[4] appears below or at this level (data not shown), namely 6σ for the most ordered part of the detergent, which constitutes an intrinsic positive control for the presence of anaesthetics. At 3.2- to 3.3-Å resolution it may not be justified to model the bound molecules individually as this will result in a different orientation in each subunit. For propofol we used fivefold NCS-averaged maps to build the molecules. For desflurane we also used NCS maps; in addition, both plausible orientations were tried and the one with the trifluoromethyl group at the bottom of the cavity was selected by comparing difference maps after refinement and independent docking scores (data not shown).

Lipids surrounding the anaesthetics binding site were partly modelled, in NCS-averaged maps. Identification of the chemical nature of the lipids is not possible with such maps at this resolution and thus we arbitrarily used phosphatidylcholine. As in the apo- protein model, the acyl chains are more ordered than the polar heads. One small part of an acyl chain corresponding to a second layer of lipid with no direct interaction with the protein is present. Structural analysis and figure preparation were done with PyMOL[36], VMD[37] and Molprobity[38]. Refinement parameters for ligands were generated with the PRODRG server[39].

Electrophysiology. Conditions for electrophysiological experiments were as follows (apart from a few variations[6] in experiments using desflurane).

Cell injection. Defolliculated, stage VI[40] *X. laevis* oocytes were obtained from a commercial supplier one day after ovary dissection. DNA (<2 ng) was blind injected into the nucleus through the animal pole, using a pneumatic microinjector, as a mixture in water of GLIC cDNA in a pmt3 vector (0.08 g l^{-1}) and green fluorescent protein (GFP) cDNA in the same vector (0.02 g l^{-1}), as a reporter gene for successful intranuclear injection. Identified cells were kept in 96-well plates with U-shaped bottom, in a HEPES-buffered modified Barth's[41] solution (in mM: NaCl 88, KCl 1, $NaHCO_3$ 2.4, HEPES 20, $MgSO_4$ 0.82, $Ca(NO_3)_2$ 0.33, $CaCl_2$ 0.41; pH 7.4; 0.22 μm filtered), at 18 °C for two days and then at 15 °C. Oocytes with the T255A mutant were transferred to pH 8 Barth's solution 2 days after injection. GFP-positive oocytes selected 2 days after injection were recorded 2–6 days after injection.

Electrophysiological recordings. Oocytes[42] were superfused with the animal pole facing a gravity driven solution inflow (4–8 ml min^{-1}) in a corridor-shaped recording chamber (flow section <10 mm^2). A solution of (in mM) NaCl 100, KCl 3, $CaCl_2$ 1, $MgCl_2$ 1 and 2-morpholino-ethanesulphonic acid (MES) 10 was adjusted to pH 8.0 with NaOH, and used as intertest on T255A oocytes. The control extracellular pH (7.3) and lower pH values were reached by adding HCl 2 mol l^{-1} and any extra pH 8.0 solution. The whole oocyte plasma membrane was voltage clamped (GeneClamp 500, Axon Instruments/Molecular Devices) using two intracellular pipettes filled with 3 M KCl (0.8–1.5 MΩ), and distinct current and voltage extracellular electrodes separately bridged to the bath near suction using 5 g l^{-1} agar in 3 M KCl. Currents were recorded (pClamp8, Axon Instruments) at air-conditioned room temperature (21–23 C), acquired at 500 Hz after low-pass filtering (200 Hz), and further filtered using 100- to 1-mean sample data reduction for figure display. Proton concentration–response curves were established at a holding potential of −50 mV, using manually controlled 30-s test-pH applications (or shorter when desensitization/inhibition at low pH produced a peak current) separated by 30- to 60-s wash at pH 7.3, or 8.0 for the T255A GLIC mutant.

Preparation of general-anaesthetic solutions. Desflurane was obtained from Baxter Healthcare Corporation. Desflurane solutions were made up gravimetrically, in gas-tight ground-glass syringes, and vigorously agitated. The final concentration was spot-checked by headspace gas chromatography. Propofol (2,6 diisopropylphenol) was obtained from Aldrich (W50,510-2). Propofol stock solution was made by dissolving pure oily liquid propofol at 1 mol l^{-1} in dimethyl sulphoxide (DMSO). It was kept in glass in the dark at room temperature for up to 1 month, and diluted in the recording solution less than 2 h before use on each cell, under strong agitation to 0.1 mM and then to lower concentrations. For most of the recordings, a single concentration of propofol was present in the perfusion system, and each cell received a single application of propofol. The perfusion system was extensively cleaned and partly replaced before going from high to low propofol concentrations.

Fit of data and statistics. Data in Fig. 3b–e are presented as mean ± s.d. Plots shown in Fig. 3b, d of log(concentration/mol l^{-1}) versus mean effect were fitted with a sigmoid function. Parameters given in Supplementary Table 2, and in Supplementary Table 3 for desflurane, were obtained by fitting for each cell a plot of concentration/mol l^{-1} versus normalized effect over four or five orders of magnitude with the Hill equation, giving values of n_H and half-maximum effective concentration (EC_{50}) (proton) or IC_{50} (desflurane) for individual cells, of which mean ± s.d. values are shown. Parameters given in Supplementary Table 3 for propofol were obtained by Hill-fitting a scatter plot of individual concentration/mol l^{-1} versus percentage inhibition data points obtained from all the cells tested (one data point per cell in most cases); n_H and IC_{50} are given with the standard error of the parameters determined from the nonlinear regression. Data in Supplementary Fig. 2 were empirically fitted with a straight line, or a simple exponential decay, to improve readability.

Molecular dynamics. We used a full atomic model of GLIC at pH 4.6 derived from previous simulations[26] to add general-anaesthetic molecules at the positions determined in the crystal structures presented in this work. The protonation state was assigned similar to previous simulations on the basis of pK_a calculations with the Yasara software[43] to represent the most probable pattern at pH 4.6, with residues E26, E35, E67, E75, E82, D86, D88, E177 and E243 being protonated. H277 was doubly protonated. The model was inserted in a fully hydrated palmitoyl-2-oleoyl-sn-glycerol-phosphatidylcholine (POPC) lipid bilayer (307 lipids, approximately 44,000 water molecules) leading to an initial system size of 128 Å × 125 Å × 182 Å. The net charge of the system was neutralized with 54 Na+ and 89 Cl$^-$ counterions, achieving a salt concentration of about 100 mM. These steps were performed within VMD[37], using the psfgen, membrane, solvate and autoionize plug-ins. Similar models were derived for the I202A, V242M and T255A mutants. The simulations were performed with NAMD[44] using the CHARMM27 (ref. 45) force field. Parameters for propofol were provided by G. Brannigan[9].

A short equilibration was performed by minimizing the system for 1,000 steps. This was followed by 30 ns of production runs. Simulations were performed at 310 K using Langevin dynamics with a damping coefficient γ of 1 ps^{-1} for temperature control. A Langevin piston algorithm was used to maintain the pressure at 1 atm. A short 10-Å cutoff was used for non-bonded interactions. A smooth switching function was used for van der Waals interactions between 8.5 and 10 Å. Long-range electrostatic interactions were treated using the particle mesh Ewald method[46]. The r-RESPA multiple time step method[47] was used with a 2-fs time step for bonded and for short-range non-bonded interactions, and a 4-fs step for long-range electrostatic forces. All bonds between hydrogen atoms and heavy atoms were constrained with the SHAKE algorithm. All molecular dynamics simulations were performed on Vargas, an IBM Regatta Power6 machine at the Institut du Développement et des Ressources en Informatique Scientifique (IDRIS) Supercomputer Center in Orsay (France).

31. Kabsch, W. Automatic processing of rotation diffraction data from crystals of 21 initially unknown symmetry and cell constants. *J. Appl. Cryst.* **26,** 795–800 (1993).
32. Collaborative Computational Project.. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
33. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).

34. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53,** 240–255 (1997).
35. Bricogne, G. *et al.* BUSTER, version 2.8.0. Cambridge, UK: Global Phasing (2009).
36. Schrödinger, L. L. C. The PyMOL molecular graphics system, version 1.3. ⟨http://www.pymol.org⟩ (2010).
37. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14,** 33–38 (1996).
38. Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35,** W375–W383 (2007).
39. Schuettelkopf, A. W. & van Aalten, D. M. *PRODRG*: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr. D* **60,** 1355–1363 (2004).
40. Dumont, J. N. Oogenesis in *Xenopus laevis* (Daudin), I. Stages of oocyte development in laboratory maintained animals. *J. Morphol.* **136,** 153–180 (1972).
41. Barth, L. G. & Barth, L. J. Differentiation of cells of the *Rana pipiens* gastrula in unconditioned medium. *J. Embryol. Exp. Morphol.* **7,** 210–222 (1959).
42. Kusano, K., Miledi, R. & Stinnakre, J. Cholinergic and catecholaminergic receptors in the *Xenopus* oocyte membrane. *J. Physiol. (Lond.)* **328,** 143–170 (1982).
43. Krieger, E., Nielsen, J. E., Spronk, C. A. & Vriend, G. Fast empirical pK_a prediction by Ewald summation. *J. Mol. Graph. Model.* **25,** 481–486 (2006).
44. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26,** 1781–1802 (2005).
45. MacKerell, A. D. Jr *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102,** 3586–3616 (1998).
46. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: an *N* log(*N*) method for Ewald sums in large systems. *J. Chem. Phys.* **98,** 10089–10092 (1993).
47. Tuckerman, M., Berne, B. J. & Martyna, G. J. Reversible multiple time scale molecular dynamics. *J. Chem. Phys.* **97,** 1990–2001 (1992).

CORRECTIONS & AMENDMENTS

ERRATUM

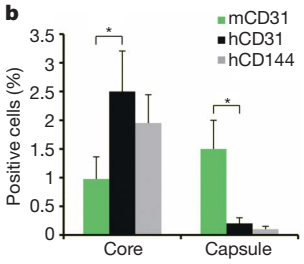
doi:10.1038/nature09734

Tumour vascularization via endothelial differentiation of glioblastoma stem-like cells

Lucia Ricci-Vitiani, Roberto Pallini, Mauro Biffoni, Matilde Todaro, Gloria Invernici, Tonia Cenci, Giulio Maira, Eugenio Agostino Parati, Giorgio Stassi, Luigi Maria Larocca & Ruggero De Maria

Nature **468,** 824–828 (2010)

In Fig. 3b of this Letter, the black bar was inadvertently labelled as mCD31 instead of hCD31 (human CD31). The corrected Fig. 3b is shown below.



CAREERS

COLUMN Procrastination can actually improve your time management **p.433**

POSTDOC JOURNALS Read about career ups, downs and prospects **go.nature.com/3fttcj**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

JUPITERIMAGES.COM

FUNDING

Researching outside the box

Open innovation offers scientists novel ways to apply their expertise — and sometimes provides much-needed cash.

BY CRISTINA JIMÉNEZ

In March 2008, Ahmet Karabulut was a couple of months from finishing a master's degree in molecular genetics, and was thinking about what to do next. During a coffee break, he read a news article about a company that posted other companies' unresolved scientific and business problems online. Anyone could send in a solution. Intrigued, Karabulut sought out the website and applied. He ended up successfully inventing a solution to an organic-chemistry problem on the stability of a compound in a common nasal decongestant. Karabulut used little more than his prior knowledge of the topic and access to the scientific literature. His immediate reward for a couple of days of work was US$20,000, crucial support as he was between jobs. But Karabulut says that the experience did more than earn him money — it also enriched his career.

He had answered the call of InnoCentive, an organization based in Waltham, Massachusetts, that uses the Internet to link 'seekers' — client companies struggling with pressing scientific or business problems — with 'solvers' — more than 250,000 problem-cracking minds around the globe, InnoCentive claims. The company is one of several, most of which are based in the United States, that are engaged in 'open innovation'. The motivating principle behind open innovation is that companies and other institutions should take full advantage of widely distributed knowledge in a wired world, finding products, patents and solutions — scientific, technological or social — outside the confines of their own organizations.

The trend has its own lingo: InnoCentive and others, including IdeaConnection of Vancouver, Canada, NineSigma of Cleveland, Ohio, TopCoder of Glastonbury, Connecticut, and yet2.com of Needham, Massachusetts, are known as innovation intermediaries, facilitators or technology brokers. The particulars of the organizations vary, but all pursue solutions to pressing research problems by posting challenges online, with the promise of financial awards — essentially crowdsourcing. Seekers range from drug to oil companies, and government agencies to non-profit organizations, and typically offer several thousand dollars for a project with a short turnaround. For scientists in need of a short-term financial boost or a supplemental source of income, open-innovation opportunities can offer a novel and challenging way to tackle new topics and add credentials to their CVs.

FRINGE BENEFITS

Since his first foray into open innovation, Karabulut has solved two further challenges, including developing a novel method for membrane-protein expression. He now works in drug discovery at the Fred Hutchinson Cancer Research Center in Seattle, Washington. "I was glad that I put the awards in my CV," he says, noting that he started to get many more responses to job applications once he had done so. Karabulut believes that the challenges showed that he could troubleshoot and think creatively, skills that many employers value. Henry Chesbrough, executive director of the Center for Open Innovation at the University of California, Berkeley, agrees. "Even those who do not 'win' often develop their ▶

reputation as an expert, making them more attractive within their innovation community for hire as a consultant or an employee for some future activity," he says. Chesbrough coined and promoted the term 'open innovation' in his book *Open Innovation: The New Imperative for Creating and Profiting from Technology* (Harvard Business School Press; 2003).

Scientists interested in becoming solvers can typically sign up for free on the intermediary's website. Usually facilitators don't require potential solvers to have formal qualifications, but InnoCentive estimates that 61% of their solvers hold an advanced degree. After signing in, candidate solvers can browse through numerous challenges in categories such as global health, clean technology or aerospace. InnoCentive, for example, runs about 40 challenges at a time. Solvers tend to choose to work on challenges that they think they can solve, says Lars Bo Jeppesen, an associate professor at Copenhagen Business School who has been studying the nature of innovation in a world of distributed knowledge for the past 10 years. This natural self-selecting method allows "the right people to solve the right problem at the right time", says Dwayne Spradlin, chief executive of InnoCentive. One major consumer goods product company claims that it has awarded prizes for more than half of its posted challenges. As is common practice, neither InnoCentive nor the company would provide specifics, fearing revealing too much to competitors.

So what does it take to become a successful solver? Some solvers emphasize the importance of a diverse research and training background and the ability to apply solutions and tools between fields or in a new one. "In fields that I haven't mastered, I can think outside of the box; I am not limited by the known and unknowns or the rules of the field," says Mounir Errami, a biochemist and bioinformatician at the University of Texas Southwestern Medical Center in Dallas. He has solved three challenges at InnoCentive, developing technologies for medicine and cosmetic chemistry unrelated to bioinformatics. Yury Bodrov, an organic chemist who has solved multiple challenges at InnoCentive and IdeaConnection, identifies two major assets for solvers. One is an analytical mind that can distinguish important data from noise. "The second is not a skill," says Bodrov, an independent consultant. "It's imagination."

INNOVATION COMPENSATION

Compensation ranges from $5,000 to $50,000, depending on the type of challenge. Sometimes, a few days' work can earn the solver several thousand dollars. At other times, however, the reward may not be worth the time and effort. Errami says that in the case of one challenge he took on, the amount of work needed has not been in line with the potential compensation. For past efforts, he says, he has typically required a short time to devise a solution, followed by a bit more to sort out technical aspects with a collaborator. "If the solution takes more than a couple of days, I avoid competing because the odds of winning are slim," he says. At InnoCentive, if a solution meets only some of the seeker's criteria, a client-services team judges how close it is. The seeker then decides whether to issue a full or partial award, or none at all.

"Even those who don't 'win' develop their reputation as an expert, making them more attractive for hire."
Henry Chesbrough

But the benefits transcend money and CV augmentation. Winning a challenge can boost a solver's confidence in his or her problem-solving skills, a benefit for those aspiring to academia or entrepreneurship. Simone Sergi, a telecommunications engineer and network system administrator at a bank in Reggio Emilia, Italy, with a doctorate in digital communications, won an InnoCentive challenge on reorganizing communications among satellites. He says that it gave him the confidence to consider a long-time dream of launching a wireless-technologies start-up company. Sergi says that he does not enjoy his job, and relished the opportunity to take on a new challenge.

Scientists who moonlight as solvers can use novel means to explore long-standing problems. Chris Wilmer, a PhD student in chemical and biological engineering at Northwestern University in Evanston, Illinois, solved a challenge that involved the lack of access to clean water in poor villages of developing countries. He found inspiration in the success of a mobile-phone business run by the Grameen Bank in Dhaka, Bangladesh. The bank would lend phones to a local entrepreneur in a poor village, who would pay back the loan by charging others in the village to use a phone. It was a profitable and self-sustaining model. Wilmer described how a similarly structured safe-water business could yield humanitarian benefits. "I enjoy solving social problems, so it was fun," says Wilmer, who has also won a second challenge involving programming a software tool to help characterize synthetic DNA strands.

For some, pursuing open innovation is a way to achieve career independence and flexibility. Grace Kepler, part-time associate research professor at the Center for Research in Scientific Computation at North Carolina State University (NCSU) in Raleigh, sees the challenges as well suited to scientists who, like her, are not employed full-time and have family responsibilities — as well as those who need more money or are unencumbered by intellectual-property issues. She applied her mathematical and statistical skills to a computational-biology challenge, and now also works as a scientific consultant, helping the company that posed the challenge to model crops. But scientists employed full- or part-time should tread carefully. When Kepler first started to work as a solver for InnoCentive, she had to obtain a waiver from NCSU stating that the university had no claim on the intellectual property in her solution, and that she had not designed it in the context of her work at the university. Seeker companies are careful to avoid revealing much to competitors — solvers often don't even know which company they're working for.

Open innovation could also offer scientists in lesser-known locations or institutions the power to break down barriers. Intermediaries solicit ideas from scientists all over the planet and from all sorts of backgrounds; an Internet connection is typically the only requirement, says Karim Lakhani, a professor of technology and operations management at Harvard Business School in Boston, Massachusetts. He calls solvers the "unusual suspects" — scientists and others with a potentially unconventional perspective.

NEW FUNDING MODEL

It is not just individuals who stand to benefit from an open-innovation funding stream; universities seeking a bit of extra money could also reap rewards. The National Physical Laboratory (NPL) in Teddington, UK, discovered open innovation while investigating alternative funding paths. "For the past three years, we've been trying to reduce our reliance on traditional government-funded routes for collaborative R&D," says Matt Smith, business development manager at the NPL, adding that the laboratory has earned more than £300,000 (US$467,000) through NineSigma, which focuses on pairing organizations — whether inventors, start-ups or universities — that possess innovative technology with companies that can commercialize those inventions. And when the National Aerospace Laboratory of the Netherlands (NRL) in Amsterdam was undergoing budget cuts, it secured a lucrative industrial research project through NineSigma to develop a life-assessment model for gas turbines. "The project lasted for 15 months and was worth several hundred thousand euros," says Arjen Vollebregt, a department manager at the NRL. He thinks that companies such as NineSigma could help labs to get extra funding with relatively little effort.

J. BLOCK

"Companies need to adapt to work with external ideas and inventors."
Wim Vanhaverbeke

Even a government agency, steeped in bureaucracy and decades of tradition, may have something to gain from open innovation — and

researchers eager to work with that agency could benefit. Bodrov solved a challenge for NASA on keeping food fresh in space, and he is now an independent scientific consultant for the chemical industry. He is also an entrepreneur: the approximately $160,000 that Bodrov earned from solving 16 challenges at InnoCentive and IdeaConnection has helped him to launch a start-up that develops nanomaterials for drug delivery in Saint Petersburg, Russia.

Jeffrey Davis, director of NASA's Space Life Sciences Programme at the Johnson Space Center in Houston, Texas, says that the agency liked Bodrov's food-packaging idea because it used a flexible graphite material — a solution perhaps familiar to materials scientists, but one that the food industry would not have generated. NASA is now looking at creating its own problem-solving framework, blending open-innovation challenges with traditional grants, contracts and small-business proposals, says Davis.

Davis says that accessing open-innovation channels was easy. "But there was a psychological barrier to admitting we couldn't find the answers ourselves," he says. This is not uncommon. Hesitation to accept outside inventions — a 'not-invented-here' stigma — is one of the major obstacles for open-innovation mechanisms, says Wim Vanhaverbeke, a professor of business studies at Hasselt University in Diepenbeek, Belgium. Companies must confront the same barrier. "They need to forget the idea that their only mission is to protect inside inventions and adapt to work with external ideas and inventors," says Vanhaverbeke.

The open-innovation approach continues to evolve. IdeaConnection, formed in 2007, is attempting to form teams of solvers using the extensive information collected in their online applications; each team member receives equal compensation if they win the challenge. Although teams composed of members with complementary backgrounds might have a better chance of solving challenges, efficiently communicating ideas between disparate members can be a challenge — something team facilitators attempt to address. InnoCentive, meanwhile, is creating a sort of 'dating site' for scientists, so that they can choose who they want to work with.

Karabulut says that the open-innovation strategy still has plenty of room to grow, "I don't know any better way for 'seekers' to find global talent for very specific challenges," he says. The thrill of winning continues to be a big part of the appeal. "It is one thing to win a cash reward," says Errami. "But it is quite a feeling to win a challenge." ■

Cristina Jiménez *is a freelance writer based in Barcelona, Spain.*

IKON IMAGES/CORBIS

COLUMN

Confessions of a procrastinator

Everyone puts off big tasks with smaller ones, and the only solution is to fight fire with fire, says **Fabio Paglieri**.

In a memorable passage from Jerome K. Jerome's 1889 novel *Three Men in a Boat*, the narrator diagnoses himself with nearly every possible ailment after leafing through a medical book found in the British Museum. Psychology researchers such as myself are prone to a special brand of hypochondria: like Jerome's character, I cannot help but wonder whether I suffer from some of the psychological shortcomings that I observe each day in the lab.

My work studying how people schedule various tasks over time (usually inefficiently) has shown me the error of my own organizational ways, and now I know the name of my terminal illness: procrastination. I am always struggling to stick to multiple deadlines on the most disparate jobs. For every project with a deadline that I manage to meet, there are two more that I am forced to postpone. I am a pathological procrastinator.

For some time I thought I was alone in my depravity, and I laboured to keep it hidden from family, friends and co-workers. Then it dawned on me: procrastination is no exotic malaise, but rather a pandemic virus, one possessed of alarming virulence in the research community. Colleagues never tire of mentioning 'bottomless to-do lists', 'overwhelming commitments', 'busy schedules' and 'pressing deadlines'. Such symptoms can result in students failing to deliver data, a co-author unable to complete a paper or a publisher postponing a manuscript's publication. Clearly I am in no position to judge, as I myself have committed similar misdeeds. I take some heart in sharing the guilt with so many others.

How might young scientists manage to avoid wrecking their careers despite such a character flaw? Procrastination often stems from over-commitment, so simply taking on fewer obligations might solve the quandary. But this is easier said than done, especially for a postdoctoral researcher. One never knows which project might turn out to be a means to new career avenues or to tenure. And by the time one realizes that a new task is just another time-consuming burden, it is often too late to retreat without repercussions.

I was about to give in to despair and start roaming the self-help aisle of my favourite bookstore in search of a cure when I found a possible solution at structuredprocrastination.com. On the site, John Perry, a professor of philosophy at Stanford University in California, notes that procrastinators are never really idle; instead, they work on something in order to put off doing something else. According to Perry, you can make procrastination work for you. Just convince yourself that there is something really complex and important that you intend to do (say, write a full monograph on your favourite research topic), and your procrastination instinct will immediately drive you to do other tasks as a way of putting off working on your big project. The trick is to make sure that these other tasks are productive and not a waste of time. The bigger your ultimate aim, the more likely you are to take part in useful procrastination chores such as running experiments, tutoring students, writing articles or going to conferences.

If Perry is right, you don't have to conquer your base procrastination impulse to progress in your professional life. True, a modicum of self-deception is required for the strategy to work. But fortunately, procrastinators are skilled self-deceivers anyway.

Will it work? It has for me so far. I have managed to diligently complete many small but important tasks as a way of putting off other impending obligations. And, unfortunately, the alternative is to conquer procrastination by sheer willpower, which is something that humans just aren't very good at. ■

Fabio Paglieri *keeps a Postdoc Journal at go.nature.com/3fttcj and is a postdoc in cognitive psychology at the Institute for Cognitive Science and Technologies of the National Research Council in Rome.*

researchers eager to work with that agency could benefit. Bodrov solved a challenge for NASA on keeping food fresh in space, and he is now an independent scientific consultant for the chemical industry. He is also an entrepreneur: the approximately $160,000 that Bodrov earned from solving 16 challenges at InnoCentive and IdeaConnection has helped him to launch a start-up that develops nanomaterials for drug delivery in Saint Petersburg, Russia.

Jeffrey Davis, director of NASA's Space Life Sciences Programme at the Johnson Space Center in Houston, Texas, says that the agency liked Bodrov's food-packaging idea because it used a flexible graphite material — a solution perhaps familiar to materials scientists, but one that the food industry would not have generated. NASA is now looking at creating its own problem-solving framework, blending open-innovation challenges with traditional grants, contracts and small-business proposals, says Davis.

Davis says that accessing open-innovation channels was easy. "But there was a psychological barrier to admitting we couldn't find the answers ourselves," he says. This is not uncommon. Hesitation to accept outside inventions — a 'not-invented-here' stigma — is one of the major obstacles for open-innovation mechanisms, says Wim Vanhaverbeke, a professor of business studies at Hasselt University in Diepenbeek, Belgium. Companies must confront the same barrier. "They need to forget the idea that their only mission is to protect inside inventions and adapt to work with external ideas and inventors," says Vanhaverbeke.

The open-innovation approach continues to evolve. IdeaConnection, formed in 2007, is attempting to form teams of solvers using the extensive information collected in their online applications; each team member receives equal compensation if they win the challenge. Although teams composed of members with complementary backgrounds might have a better chance of solving challenges, efficiently communicating ideas between disparate members can be a challenge — something team facilitators attempt to address. InnoCentive, meanwhile, is creating a sort of 'dating site' for scientists, so that they can choose who they want to work with.

Karabulut says that the open-innovation strategy still has plenty of room to grow, "I don't know any better way for 'seekers' to find global talent for very specific challenges," he says. The thrill of winning continues to be a big part of the appeal. "It is one thing to win a cash reward," says Errami. "But it is quite a feeling to win a challenge." ■

Cristina Jiménez *is a freelance writer based in Barcelona, Spain.*

IKON IMAGES/CORBIS

COLUMN

Confessions of a procrastinator

Everyone puts off big tasks with smaller ones, and the only solution is to fight fire with fire, says **Fabio Paglieri**.

In a memorable passage from Jerome K. Jerome's 1889 novel *Three Men in a Boat*, the narrator diagnoses himself with nearly every possible ailment after leafing through a medical book found in the British Museum. Psychology researchers such as myself are prone to a special brand of hypochondria: like Jerome's character, I cannot help but wonder whether I suffer from some of the psychological shortcomings that I observe each day in the lab.

My work studying how people schedule various tasks over time (usually inefficiently) has shown me the error of my own organizational ways, and now I know the name of my terminal illness: procrastination. I am always struggling to stick to multiple deadlines on the most disparate jobs. For every project with a deadline that I manage to meet, there are two more that I am forced to postpone. I am a pathological procrastinator.

For some time I thought I was alone in my depravity, and I laboured to keep it hidden from family, friends and co-workers. Then it dawned on me: procrastination is no exotic malaise, but rather a pandemic virus, one possessed of alarming virulence in the research community. Colleagues never tire of mentioning 'bottomless to-do lists', 'overwhelming commitments', 'busy schedules' and 'pressing deadlines'. Such symptoms can result in students failing to deliver data, a co-author unable to complete a paper or a publisher postponing a manuscript's publication. Clearly I am in no position to judge, as I myself have committed similar misdeeds. I take some heart in sharing the guilt with so many others.

How might young scientists manage to avoid wrecking their careers despite such a character flaw? Procrastination often stems from over-commitment, so simply taking on fewer obligations might solve the quandary. But this is easier said than done, especially for a postdoctoral researcher. One never knows which project might turn out to be a means to new career avenues or to tenure. And by the time one realizes that a new task is just another time-consuming burden, it is often too late to retreat without repercussions.

I was about to give in to despair and start roaming the self-help aisle of my favourite bookstore in search of a cure when I found a possible solution at structuredprocrastination.com. On the site, John Perry, a professor of philosophy at Stanford University in California, notes that procrastinators are never really idle; instead, they work on something in order to put off doing something else. According to Perry, you can make procrastination work for you. Just convince yourself that there is something really complex and important that you intend to do (say, write a full monograph on your favourite research topic), and your procrastination instinct will immediately drive you to do other tasks as a way of putting off working on your big project. The trick is to make sure that these other tasks are productive and not a waste of time. The bigger your ultimate aim, the more likely you are to take part in useful procrastination chores such as running experiments, tutoring students, writing articles or going to conferences.

If Perry is right, you don't have to conquer your base procrastination impulse to progress in your professional life. True, a modicum of self-deception is required for the strategy to work. But fortunately, procrastinators are skilled self-deceivers anyway.

Will it work? It has for me so far. I have managed to diligently complete many small but important tasks as a way of putting off other impending obligations. And, unfortunately, the alternative is to conquer procrastination by sheer willpower, which is something that humans just aren't very good at. ■

Fabio Paglieri *keeps a Postdoc Journal at go.nature.com/3fttcj and is a postdoc in cognitive psychology at the Institute for Cognitive Science and Technologies of the National Research Council in Rome.*

LAST OF THE GUERRILLA GARDENERS

Seeding a revolution.

BY DAVID L. CLEMENTS

They came for 'Percy Thrower' last night. I was on my way to deliver some Pink Brandywine tomato seeds when I saw the first police car. I turned the corner and saw a fleet of them parked outside her house, complete with sniffer dogs and a space-suited forensic team heading for her potting shed.

I averted my eyes and walked past on the opposite side of the road, feeling the envelope of illegal seeds in my pack broadcast my guilt. As I left her road, the sterilization van arrived, its flame throwers ready to destroy 'Percy's' irreplaceable collection of plants.

I got away. The others weren't so lucky. As I waited for the bus I checked our secure server and realized they were rolling up the whole network. 'Monty' had been the first, but in catapulting a package of herb seeds into Buckingham Palace gardens he'd gone too far. His arrest had been the trigger for raids across the country. 'Bob' had sent out a warning as they smashed down his door, but they'd been ready for us all. If I hadn't been on a delivery run they'd've caught me as well.

I couldn't go home. Most of the people I trusted had been picked up. I stayed on the bus as it passed my stop and headed into central London. The clean-up crews were obvious, torching collections of wild flowers in the roadside beds that I'd seeded from bus windows while commuting.

All the hard work, all the beautiful, irreplaceable diversity, stamped out by commercial greed. If I'd've had the machinery with me I'd've leapt off the bus and seeded the palace gardens myself.

'Percy' had started the whole thing with a few prophetic words: "Biology is the biggest peer-to-peer copying system on the planet. Now they've eliminated file sharing they'll come for the seed sharers."

She'd been a university botanist for years but left when it became clear that all the grants were controlled by big agribusiness. We knew we were in trouble when Kew was sold off and Henry Doubleday broken up. Their vast seed collections became the intellectual property of a few huge corporations. Unlicensed seeds were already illegal to sell, but once companies owned the rare strains, they stopped collectors sharing them for free. They wanted to control it all.

At first we tried to stop them. There were protests, lobbies and mass marches. *Gardener's Question Time* became such a political hot potato it was cancelled by the BBC. And then came the Chelsea Flower-Show Riots.

When I got off the bus I saw the police at the station. But it was just the usual patrols, not yet a manhunt. Maybe word of my escape had yet to reach them. I headed for Left Luggage.

We were called economic terrorists, threatening profits from high-cost, high-yield, terminator-gene strains that would feed the world and soak up excess CO_2. But we just wanted tasty vegetables from our own gardens, unusual flowers smelling as good as they looked, and the opportunity to eat the occasional purple carrot. Serious action only came when self-propagating super-plants were found growing by a road in Norfolk.

"Businessmen don't understand that biology is a lot messier than digital copying," 'Percy' had said as we talked in her potting shed. "There's a dozen perfectly natural ways the terminator gene might have failed. One cosmic ray taking out the right base pair would be enough!" But scientific sense was never going to stand up to irate politicians shouting "Something must be done!" Fines became prison sentences, the Seed Squads were established and we were forced underground. Home gardens were no longer safe, so we became guerrilla gardeners — a secret society sharing seeds and planting contraband crops in public spaces. We tended them at night or just scattered seeds far and wide to let nature take its course. That's when the network started and we adopted our *noms de vert*.

➲ NATURE.COM
Follow Futures on Facebook at:
go.nature.com/mtoodm

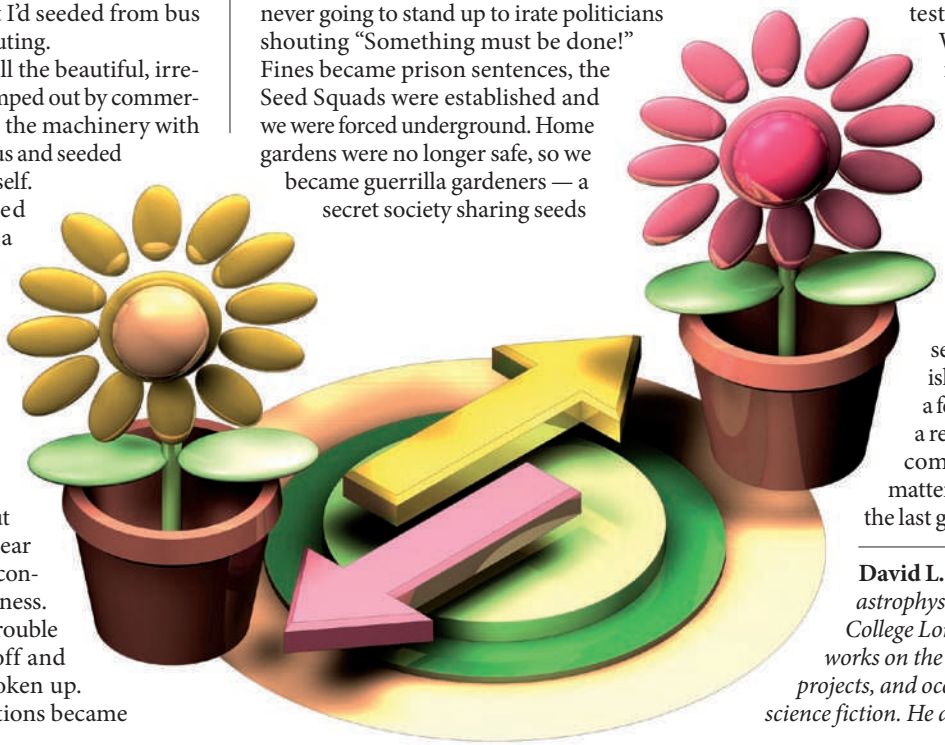
We were too successful. Nature was indeed the great copier. Our wilder strains could fend for themselves and started to spread. The gloves finally came off when the director of Smaxo's agricultural division found a clump of illegal Afghan Purple carrots growing at the bottom of his garden and carpeted the Prime Minister. Of course, 'Monty' and his catapult didn't help.

Now the only guerrilla gardener left is me.

I collected my escape stash from Left Luggage along with Monty's seed catapult. The wig, hat and glasses helped me slip past the tighter police patrols and onto the sleeper to Fort William. Locked in my cabin I shaved my distinctive beard and used clippers so that I'd match the fake ID in my stash. The train would go through a lot of isolated country. The clean up crews couldn't cover all that ground in one season, so some of my seeds were going out of the window.

As for the rest …

There are islands off the Scottish coast contaminated by bioweapons testing from the Second World War. People are forbidden and there are no sheep or rabbits. But their climate is ideal. The seed catapult has enough range to reach them from a boat offshore. Or I could land and make certain they're properly planted. The seeds will do well on the islands, even if I don't. In a few years they'll become a reserve for natural, non-commercial diversity no matter what happens to me, the last guerrilla gardener. ■



JACEY

David L. Clements *is an astrophysicist at Imperial College London where he works on the Herschel and Planck projects, and occasionally writes science fiction. He doesn't have a garden.*

LETTER

doi:10.1038/nature09638

Structure of human *O*-GlcNAc transferase and its complex with a peptide substrate

Michael B. Lazarus[1,4]*, Yunsun Nam[2,3]*, Jiaoyang Jiang[4], Piotr Sliz[2,3] & Suzanne Walker[4]

The essential mammalian enzyme O-linked β-*N*-acetylglucosamine transferase (*O*-GlcNAc transferase, here OGT) couples metabolic status to the regulation of a wide variety of cellular signalling pathways by acting as a nutrient sensor[1]. OGT catalyses the transfer of *N*-acetylglucosamine from UDP-*N*-acetylglucosamine (UDP-GlcNAc) to serines and threonines of cytoplasmic, nuclear and mitochondrial proteins[2,3], including numerous transcription factors[4], tumour suppressors, kinases[5], phosphatases[1] and histone-modifying proteins[6]. Aberrant glycosylation by OGT has been linked to insulin resistance[7], diabetic complications[8], cancer[9] and neurodegenerative diseases including Alzheimer's[10]. Despite the importance of OGT, the details of how it recognizes and glycosylates its protein substrates are largely unknown. We report here two crystal structures of human OGT, as a binary complex with UDP (2.8 Å resolution) and as a ternary complex with UDP and a peptide substrate (1.95 Å). The structures provide clues to the enzyme mechanism, show how OGT recognizes target peptide sequences, and reveal the fold of the unique domain between the two halves of the catalytic region. This information will accelerate the rational design of biological experiments to investigate OGT's functions; it will also help the design of inhibitors for use as cellular probes and help to assess its potential as a therapeutic target.

The ability to sense and respond to nutrient levels is critical for the growth of all living systems. In eukaryotes, a major mechanism for nutrient sensing involves the essential[11] protein glycosyltransferase OGT, which senses cellular glucose levels via UDP-GlcNAc concentrations, and responds by dynamically *O*-GlcNAcylating a wide range of nuclear and cytoplasmic proteins[1,12]. These include proteins involved in insulin-like signalling pathways[7] and transcriptional activators that regulate glucose levels by controlling gluconeogenesis[13]. As many known *O*-GlcNAcylation sites are also phosphorylation sites, OGT is proposed to play a major role in modulating cellular kinase signalling cascades[14]. OGT is also involved in widespread transcriptional regulation[15–17]. Prolonged hyperglycaemia, such as occurs in diabetes, or excessive glucose uptake, such as occurs in cancer cells, results in hyper-*O*-GlcNAcylation of cellular proteins by OGT, and this increased *O*-GlcNAcylation has been linked to harmful cellular effects[18]. Thus, strategies to modulate OGT activity may have therapeutic value for treating diabetic complications, cancer, and other diseases[13].

The lack of a crystal structure has been a major impediment to investigating OGT's molecular mechanisms, understanding substrate recognition, and developing inhibitors. OGT comprises two distinct regions: an N-terminal region consisting of a series of tetratricopeptide repeat (TPR) units[19,20] and a multidomain catalytic region. The TPR domain is proposed to scaffold interactions with other proteins, which may play a role in determining substrate selectivity[21]. A crystal structure comprising 11.5 TPR units of human OGT has been reported[21], but there have been no structures of the catalytic region. From sequence analysis and structures of bacterial glycosyltransferases[22–26],

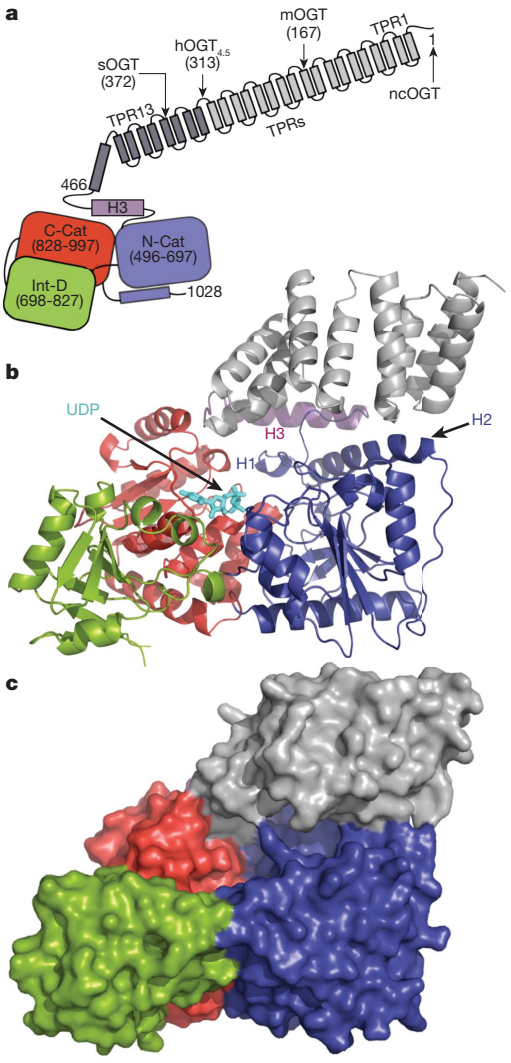


Figure 1 | Overall structure of human OGT complexed to UDP. **a**, Schematic of OGT domain architecture with the TPR units shown in grey, the transitional helix (H3) in purple, the N-Cat domain in blue, the Int-D domain in green, and the C-Cat domain in red. The native isoforms of OGT (sOGT, short OGT; mOGT, mitochondrial OGT; and ncOGT, nucleocytoplasmic OGT) and the crystallization construct differ only in the number of TPRs, as shown. **b**, Overall fold of OGT from the OGT–UDP complex in a ribbon representation. The colouring is the same as the schematic in **a**. The UDP is shown in cyan. The N-Cat domain helices unique to OGT are indicated as H1 and H2. **c**, Surface representation of the OGT–UDP complex. The colouring scheme is the same as in **a** and **b**.

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [2]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA. [3]Laboratory of Molecular Medicine, Children's Hospital, Boston, Massachusetts 02115, USA. [4]Department of Microbiology and Molecular Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.
*These authors contributed equally to this work.

including a bacterial homologue of unknown function[25,26], OGT was predicted to be a member of the GT-B superfamily of glycosyltransferases (Gtfs)[27]. However, OGT is unusual because it is the only known member to glycosylate polypeptides and it contains a long uncharacterized intervening sequence (~120 amino acids) in the middle of the catalytic region. It is also proposed to contain a phosphatidylinositol (3,4,5)-trisphosphate (PIP_3) binding domain involved in membrane recruitment in response to insulin signalling[7].

We report two crystal structures of a human OGT construct ($hOGT_{4.5}$) containing 4.5 TPR units and the catalytic domain. The catalytic properties of this construct are similar to those of the full-length enzyme (Supplementary Fig. 1)[28]. One structure (2.8 Å, referred to as OGT–UDP) is a complex with UDP; the other structure (1.95 Å, referred to as OGT–UDP–peptide) is a complex containing UDP and a well-characterized 14-residue CKII peptide substrate[28]. On the basis of currently available experimental data, we also present a model for the full-length enzyme (Supplementary Information). Details of structure determination are presented in Methods and Supplementary Tables 1 and 2.

The OGT–UDP complex is shown in Fig. 1. The catalytic region contains three domains: the amino (N)-terminal domain (N-Cat), the carboxy (C)-terminal domain (C-Cat), and the intervening domain (Int-D) (Fig. 1a, b). The N-Cat and C-Cat domains have Rossmann-like folds typical of GT-B superfamily members; however, the N-Cat domain is distinctive in containing two additional helices, H1 and H2, which form an essential part of the active site (Fig. 1b). The Int-D domain, which has a novel fold, packs exclusively against the C-Cat domain (Fig. 1c). The UDP moiety binds in a pocket in the C-Cat domain near the interface with the N-Cat domain[27]. This pocket is lined with conserved residues shown to be important for catalytic activity (Supplementary Table 3)[25,26]. A transitional helix (H3) links the catalytic region to the TPR repeats, which spiral along the upper surface of the catalytic region from the C-Cat domain to the N-Cat domain. The TPRs and the catalytic region are demarcated by a narrow horizontal cleft.

The OGT–UDP–peptide complex (Fig. 2), which crystallized in a different space group from the OGT–UDP complex, has a wider cleft between the TPR domain and the catalytic region than the OGT–UDP complex (Fig. 1c and Fig. 2a), and the CKII peptide binds in this cleft. This peptide, YPGGSTPVS*SANMM, contains three serines and a threonine, but only one serine (underlined; referred to as Ser*) is glycosylated by OGT[28]. The hydroxyl of Ser* points into the nucleotide-sugar binding site (Fig. 2b). The two residues N-terminal to Ser* lie over the UDP moiety; the residues C-terminal to Ser* traverse towards the back of the cleft along the H2 helix of the N-Cat lobe. Although OGT glycosylates a wide range of target peptides, it prefers sequences in which the residues flanking the glycosylated amino acid enforce an extended conformation (for example, prolines and β-branched amino acids; see Supplementary Fig. 2 and Supplementary Table 4). Consistent with these preferences, the peptide is anchored mainly by contacts from

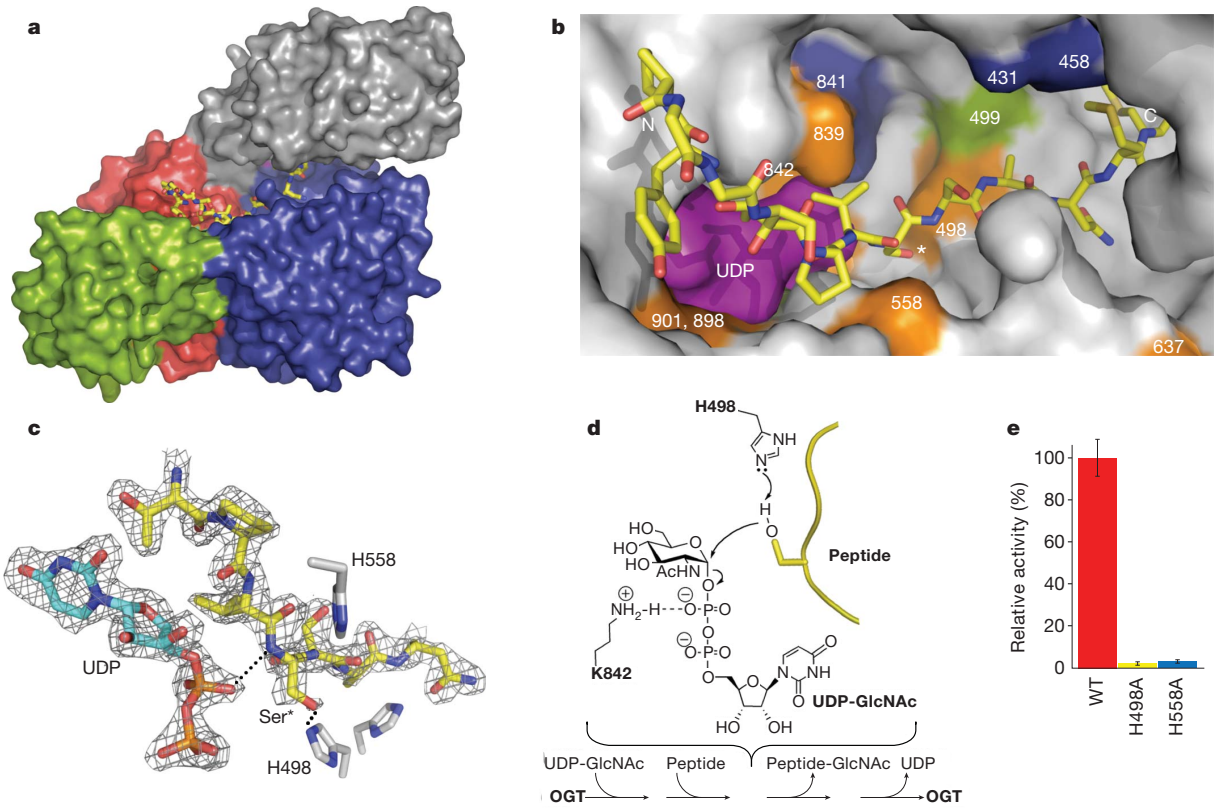


Figure 2 | Structure of the OGT–UDP–peptide complex. **a**, Surface rendering of the OGT complex with UDP and the CKII peptide substrate[28]. The view and the colouring is the same as in Fig. 1. The peptide, shown in yellow, lies over the UDP moiety, which is not visible in this orientation. **b**, Close-up surface rendering of the OGT active site (grey) containing the CKII peptide in a stick representation (carbon atoms shown in yellow) with the UDP (purple) in a space filling representation lying directly underneath it. The reactive serine is indicated by an asterisk. The peptide binds in the cleft between the TPR region and the catalytic region, and extends along the interface between the C-Cat and N-Cat domains. Protein residues implicated in catalytic activity are coloured orange, green, or blue in decreasing order of importance based on residual activity after mutation (Supplementary Table 3). Lysine 842 (orange) lies underneath UDP in this view. **c**, View of UDP (carbon atoms shown in cyan) and part of the CKII peptide (carbon atoms shown in yellow) with selected OGT side chains shown. Dashed lines indicate inferred hydrogen bonds based on distances in the OGT–UDP–peptide complex. The $2F_o - F_c$ omit map is contoured at the 1σ level. **d**, Proposed mechanism of OGT. The ordered sequential bi-bi kinetic mechanism shown is based on the structure of the ternary complex and supporting kinetic experiments (Supplementary Fig. 4). The peptide is depicted in yellow with only the reactive serine hydroxyl shown. H498 is the proposed catalytic base. Lys 842, also shown to be essential for activity[25,26], stabilizes the UDP moiety. **e**, Histogram showing the relative activities of the H498A and H558A mutants compared to the wild-type (WT) protein (average ± s.d., $n = 3$).

OGT side chains to the amide backbone, with an additional contact from the UDP moiety to the backbone amide of Ser*. The cleft is also filled with ordered water molecules, enabling it to serve as an adaptable interface to bind a range of polypeptides containing side chains of different sizes, polarity, and hydrogen bonding capabilities. As the peptide substrate is anchored by contacts to its backbone, it is reasonable to infer that protein substrates are glycosylated on flexible regions such as loops or termini that can bind in an extended conformation, exposing the amide backbone.

The closed conformation of the substrate-binding cleft in the OGT–UDP structure is stabilized by a 'latch' comprising contacts between TPRs 10/11 and the H2 helix of the catalytic domain (Fig. 2a and Supplementary Fig. 3). Opening of the cleft in the OGT–UDP–peptide complex occurs owing to a hinge-like motion around a pivot point between TPRs 12 and 13. The two structures suggest that glycosylation substrates enter the active site from the face of the enzyme shown in Fig. 2a, with the TPR domain restricting or allowing access, depending on its conformation and its interactions with the catalytic domain. Molecular dynamics simulations indicate that the 'hinge' between the catalytic domain and the TPR domain is capable of large motions that fully expose the active site, which would allow protein substrates to approach closely enough for surface loops to enter (Supplementary Movie 1). The molecular mechanisms that facilitate or stabilize opening of the cleft to allow access of protein substrates remain to be determined, but may involve interactions between protein substrates or adaptor proteins and the other regions of OGT.

The OGT–UDP–peptide complex, in addition to revealing how peptide substrates bind, provides unexpected insights into the kinetic mechanism. OGT was previously proposed to have a random sequential 'bi-bi' mechanism in which either substrate can bind first[28]. The structure, however, indicates that the peptide substrate binds over the nucleotide-sugar binding pocket, blocking access to it. Moreover, the α-phosphate of the UDP moiety contacts the backbone amide of Ser* (Fig. 2c), which helps orient the peptide. The peptide complex suggests an ordered mechanism in which UDP-GlcNAc binds before the polypeptide substrate. To assess the order of substrate binding, we analysed the product inhibition patterns for UDP. At saturating peptide concentrations, a competitive inhibition pattern was obtained for UDP with respect to UDP-GlcNAc, which is inconsistent with a random mechanism, but supports the ordered sequential bi-bi mechanism implied by the crystal structure (Supplementary Fig. 4).

Another insight from the crystal structure is the identity of the catalytic base. On the basis of analyses of other GT-B family members, including the bacterial OGT homologue, it was proposed that His 558 is the catalytic base. Although we have verified that this residue is critical for catalytic activity, the peptide complex shows that it is more than 5 Å away from the reactive serine hydroxyl and makes an apparent hydrogen bond with the backbone carbonyl of the preceding residue. In contrast, His 498, which is invariant in metazoan OGTs but absent in the homologous bacterial enzyme, protrudes from helix H1 into the active site within 3.5 Å of the Ser* hydroxyl. As His 498 is critical for activity and is located between the reactive serine hydroxyl and the GlcNAc binding pocket, it is the probable catalytic base in OGT.

We were unable to obtain a crystal of the OGT–UDP-GlcNAc complex owing to hydrolysis of the substrate, but according to the computational docking experiments we performed, the GlcNAc is oriented in a manner that exposes its β-face to the overlying peptide (Supplementary Fig. 5) and places the anomeric carbon near the reactive serine. This conformation is similar to the UDP-GlcNAc conformation observed in a complex of another GT-B family member[23], and its relevance is supported by evidence that the C2 N-acetyl moiety projects up from the OGT sugar binding pocket[29]. Furthermore, it is consistent with the enzymatic reaction, which

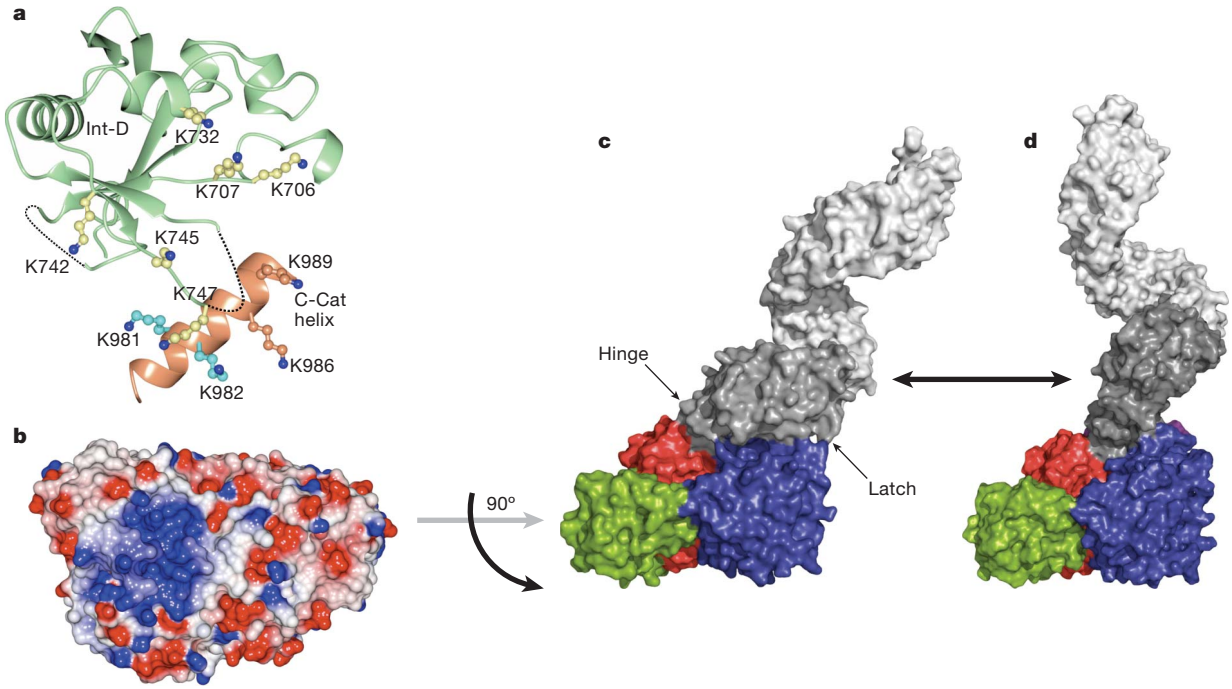


Figure 3 | Structure of the intervening domain and full-length models of human OGT. a, Ribbon representation of the intervening domain rendered in light green with missing loops represented by dotted lines. Lysine side chains that form an extensive positive surface (see **b**) are displayed in a 'ball-and-stick' representation. Shown in coral is a helix from the C-cat domain containing four basic residues that contribute to the positively charged surface[7]. **b**, Surface representation of OGT coloured according to electrostatic potential, with blue representing areas of positive charge and red representing areas of negative charge. The protein is rotated 90° around the *x*-axis from the representation shown in Figs 1, 2 and 3c, exposing the bottom surface of the catalytic region. **c**, Model of full-length human OGT, shown as a surface rendering and coloured as in Fig. 1a, based on the $hOGT_{4.5}$ structures and the previously reported TPR domain structure. The TPRs preceding the boundary of $hOGT_{4.5}$ are shown in light grey. The model is shown as a monomer, but OGT may exist in different oligomerization states in cells[21,28]. Hinge and latch regions are indicated by arrows. **d**, Model of full-length OGT opening to accommodate larger substrates. The 'open' conformation is based on molecular dynamics simulations (Supplementary Movie 1), as described in Methods. (Atomic coordinates for full-length models are available for download; see Supplementary Information.)

involves displacement of the α-UDP group to yield an inverted product. On the basis of the accumulated biochemical and structural data, we propose a general mechanism for the reaction (Fig. 2d).

The most unusual feature of OGT is the intervening domain between the catalytic lobes, which is only found in metazoans (Supplementary Figs 6 and 7). This polypeptide adopts a topologically novel fold with a seven-stranded β-sheet core stabilized by flanking α-helices (Fig. 3a). There are two long unstructured loops for which electron density is missing. An electrostatic surface rendering shows that the intervening domain and an adjacent helix of the C-Cat domain form a large basic surface comprising ten lysine residues (Fig. 3a and b). Among these are K981 and K982, which were previously reported to constitute part of a PIP_3 binding motif that recruits OGT to membranes[7]. We mutated eight of these ten lysines in various combinations (Supplementary Table 3). All mutants were catalytically active (Supplementary Fig. 8), but we were unable to identify a role for the Int-D domain in PIP_3 binding (Supplementary Table 5). We suggest that this domain is involved in other functions *in vivo*. These functions may include substrate selection, cellular localization, or interactions with regulatory factors or receptors. The reported structures and mutant data provide a crucial starting point for investigating the possible roles of the intervening domain.

The structures reported here show how OGT recognizes peptide sequences and provide new information on the enzymatic mechanism as well as a view of the intervening domain. Models of full-length human OGT in its open and closed states, constructed on the basis of crystal structures and molecular dynamics simulations, highlight the conformational changes that may regulate access of substrates to the active site (Fig. 3c and d). Our structures may assist in the development of inhibitors with possible therapeutic value for treating diseases associated with excessive *O*-GlcNAcylation.

METHODS SUMMARY

Human OGT residues 313–1031 (CPTH…KPVE) were expressed in *Escherichia coli* and purified by nickel affinity chromatography and gel filtration. Protein was then incubated with UDP or with UDP and a 17-residue substrate peptide (KKKYPGGSTPVSSANMM), which was cleaved to YPGGSTPVSSANMM in the crystallization drop (confirmed by mass spectrometry). The OGT–UDP structure was determined using the method of multiple isomorphous replacement with anomalous scattering (MIRAS) (Supplementary Table 2). The OGT–UDP–peptide complex structure was solved by molecular replacement using the refined OGT–UDP structure. The crystal packing for the two complexes is described in Supplementary Fig. 9. Kinetic analysis was performed using UDP-^{14}C-GlcNAc and a lysine tagged CKII peptide using our previously described filter binding assay[29]. The molecular dynamics simulation was performed by using the program Desmond[30] on an optimized 64-node Linux-based InfiniBand cluster.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 20 April; accepted 3 November 2010.
Published online 16 January 2011.

1. Hart, G. W., Housley, M. P. & Slawson, C. Cycling of O-linked β-N-acetylglucosamine on nucleocytoplasmic proteins. *Nature* **446,** 1017–1022 (2007).
2. Torres, C. R. & Hart, G. W. Topography and polypeptide distribution of terminal N-acetylglucosamine residues on the surfaces of intact lymphocytes. Evidence for O-linked GlcNAc. *J. Biol. Chem.* **259,** 3308–3317 (1984).
3. Haltiwanger, R. S., Holt, G. D. & Hart, G. W. Enzymatic addition of O-GlcNAc to nuclear and cytoplasmic proteins. Identification of a uridine diphospho-N-acetylglucosamine:peptide beta-N-acetylglucosaminyltransferase. *J. Biol. Chem.* **265,** 2563–2568 (1990).
4. Yang, X., Zhang, F. & Kudlow, J. E. Recruitment of O-GlcNAc transferase to promoters by corepressor mSin3A: coupling protein O-GlcNAcylation to transcriptional repression. *Cell* **110,** 69–80 (2002).
5. Dias, W. B., Cheung, W. D., Wang, Z. & Hart, G. W. Regulation of calcium/calmodulin-dependent kinase IV by O-GlcNAc modification. *J. Biol. Chem.* **284,** 21327–21337 (2009).
6. Fujiki, R. *et al.* GlcNAcylation of a histone methyltransferase in retinoic-acid-induced granulopoiesis. *Nature* **459,** 455–459 (2009).
7. Yang, X. *et al.* Phosphoinositide signalling links O-GlcNAc transferase to insulin resistance. *Nature* **451,** 964–969 (2008).
8. Brownlee, M. Biochemistry and molecular cell biology of diabetic complications. *Nature* **414,** 813–820 (2001).
9. Caldwell, S. A. *et al.* Nutrient sensor O-GlcNAc transferase regulates breast cancer tumorigenesis through targeting of the oncogenic transcription factor FoxM1. *Oncogene* **29,** 2831–2842 (2010).
10. Liu, F., Iqbal, K., Grundke-Iqbal, I., Hart, G. W. & Gong, C. X. O-GlcNAcylation regulates phosphorylation of tau: a mechanism involved in Alzheimer's disease. *Proc. Natl Acad. Sci. USA* **101,** 10804–10809 (2004).
11. Shafi, R. *et al.* The O-GlcNAc transferase gene resides on the X chromosome and is essential for embryonic stem cell viability and mouse ontogeny. *Proc. Natl Acad. Sci. USA* **97,** 5735–5739 (2000).
12. Love, D. C. & Hanover, J. A. The hexosamine signaling pathway: deciphering the "O-GlcNAc code". *Sci. STKE* **2005,** re13 (2005).
13. Dentin, R., Hedrick, S., Xie, J., Yates, J. III & Montminy, M. Hepatic glucose sensing via the CREB coactivator CRTC2. *Science* **319,** 1402–1405 (2008).
14. Wells, L., Vosseller, K. & Hart, G. W. Glycosylation of nucleocytoplasmic proteins: signal transduction and O-GlcNAc. *Science* **291,** 2376–2378 (2001).
15. Gambetta, M. C., Oktaba, K. & Muller, J. Essential role of the glycosyltransferase sxc/Ogt in polycomb repression. *Science* **325,** 93–96 (2009).
16. Sinclair, D. A. *et al. Drosophila* O-GlcNAc transferase (OGT) is encoded by the Polycomb group (PcG) gene, super sex combs (sxc). *Proc. Natl Acad. Sci. USA* **106,** 13427–13432 (2009).
17. Love, D. C. *et al.* Dynamic O-GlcNAc cycling at promoters of *Caenorhabditis elegans* genes regulating longevity, stress, and immunity. *Proc. Natl Acad. Sci. USA* **107,** 7413–7418 (2010).
18. Goldberg, H. J., Whiteside, C. I., Hart, G. W. & Fantus, I. G. Posttranslational, reversible O-glycosylation is stimulated by high glucose and mediates plasminogen activator inhibitor-1 gene expression and Sp1 transcriptional activity in glomerular mesangial cells. *Endocrinology* **147,** 222–231 (2006).
19. Kreppel, L. K., Blomberg, M. A. & Hart, G. W. Dynamic glycosylation of nuclear and cytosolic proteins. Cloning and characterization of a unique O-GlcNAc transferase with multiple tetratricopeptide repeats. *J. Biol. Chem.* **272,** 9308–9315 (1997).
20. Lubas, W. A., Frank, D. W., Krause, M. & Hanover, J. A. O-Linked GlcNAc transferase is a conserved nucleocytoplasmic protein containing tetratricopeptide repeats. *J. Biol. Chem.* **272,** 9316–9324 (1997).
21. Jinek, M. *et al.* The superhelical TPR-repeat domain of O-linked GlcNAc transferase exhibits structural similarities to importin α. *Nature Struct. Mol. Biol.* **11,** 1001–1007 (2004).
22. Ha, S., Walker, D., Shi, Y. & Walker, S. The 1.9 A crystal structure of *Escherichia coli* MurG, a membrane-associated glycosyltransferase involved in peptidoglycan biosynthesis. *Protein Sci.* **9,** 1045–1052 (2000).
23. Hu, Y. *et al.* Crystal structure of the MurG:UDP-GlcNAc complex reveals common structural principles of a superfamily of glycosyltransferases. *Proc. Natl Acad. Sci. USA* **100,** 845–849 (2003).
24. Wrabl, J. O. & Grishin, N. V. Homology between O-linked GlcNAc transferases and proteins of the glycogen phosphorylase superfamily. *J. Mol. Biol.* **314,** 365–374 (2001).
25. Martinez-Fleites, C. *et al.* Structure of an O-GlcNAc transferase homolog provides insight into intracellular glycosylation. *Nature Struct. Mol. Biol.* **15,** 764–765 (2008).
26. Clarke, A. J. *et al.* Structural insights into mechanism and specificity of O-GlcNAc transferase. *EMBO J.* **27,** 2780–2788 (2008).
27. Lairson, L. L., Henrissat, B., Davies, G. J. & Withers, S. G. Glycosyltransferases: structures, functions, and mechanisms. *Annu. Rev. Biochem.* **77,** 521–555 (2008).
28. Kreppel, L. K. & Hart, G. W. Regulation of a cytosolic and nuclear O-GlcNAc transferase. Role of the tetratricopeptide repeats. *J. Biol. Chem.* **274,** 32015–32022 (1999).
29. Gross, B. J., Kraybill, B. C. & Walker, S. Discovery of O-GlcNAc transferase inhibitors. *J. Am. Chem. Soc.* **127,** 14588–14589 (2005).
30. Bowers, K. J. *et al.* Scalable algorithms for molecular dynamics simulations on commodity clusters. *Proc. ACM/IEEE Conf. on Supercomputing (SC06)* (ACM Press, 2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank B. Gross and C. Drennan for advice. We also thank the US National Institutes of Health, the US National Science Foundation, and the Harvard Biomedical Accelerator Fund for financial support. This work is based on research conducted at the Advanced Photon Source (Northeastern Collaborative Access Team beamlines) and Brookhaven National Laboratory (X25 and X29 beamlines).

Author Contributions S.W. conceived the project. M.B.L. obtained the crystallization construct and initial diffracting crystals. M.B.L., Y.N. and P.S. determined and refined the crystal structures. J.J. and M.B.L. performed the enzymatic assays. M.B.L., Y.N., J.J., P.S. and S.W. designed experiments, discussed results, and prepared the manuscript.

Author Information The structures of the OGT–UDP complex and the OGT–UDP–peptide complex have been submitted to the Protein Data Bank under accession numbers 3PE3 and 3PE4. Atomic coordinates for the full-length models of OGT as well as the docked UDP-GlcNAc structure are available for download from the Walker Laboratory website (see Supplementary Information). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.W. (enzymology; suzanne_walker@hms.harvard.edu) or P.S. (structural biology; piotr_sliz@hms.harvard.edu).

METHODS

Protein purification. Full length human OGT (ncOGT) was expressed as previously described. The $OGT_{4.5}$ construct (spanning residues 313–1031 based on the numbering of the full length human protein) was constructed from our previously reported *E. coli* codon-optimized construct using primers listed in Supplementary Table 6 after being cloned into a pET24b vector (Novagen)[29]. After plasmid transformation into BL21 (DE3), the protein was expressed as a fusion protein with an N terminus consisting of a T7 tag, followed by an 8-His tag, followed by an HRV3C protease cleavage site (LEVLFQGP). Cultures were grown at 37 °C after diluting an overnight culture 1 to 100 in fresh LB media. Cells were grown to an A_{600} of 1.1, at which point they were transferred to a temperature of 16 °C. After letting the cells grow at 16 °C for 30 min, they were induced with 0.2 mM IPTG and grown overnight at 16 °C for 16 h. Cells were pelleted, resuspended in TBS (20 mM Tris, pH 7.4, 250 mM NaCl) supplemented with 1 mM PMSF and 0.1 mg ml^{-1} lysozyme, lysed and the lysate was centrifuged at 5,000*g* for 20 min to remove unbroken cells. The supernatant was then centrifuged at 100,000*g* to further clean the lysate. Imidazole was then added to the supernatant to a final concentration of 40 mM before the lysate was incubated with Ni-NTA agarose superflow resin (Qiagen) which was prewashed with TBS + 40 mM imidazole for batch nickel affinity purification. After incubating the lysate and the resin with gentle rocking at 4 °C, the flowthrough was removed and the resin was washed with 10 column volumes of TBS + 50 mM imidazole. The protein was then eluted with 4 column volumes of TBS + 250 mM imidazole. The eluate was supplemented with 0.5 μM THP to prevent aggregation and then concentrated with centrifugal concentrators (Millipore). After protein concentration determination, the N-terminal tags were cleaved by adding HRV 3C protease (EMD) to the concentrated purified protein at a ratio 1 unit/150 μg of protein and incubating at 4 °C for 16 h. Following cleavage, the protein was further purified by gel filtration on a Superdex 200 column (GE Healthcare) in TBS (20 mM Tris, pH 8.0, 150 mM NaCl) + 0.5 μM THP (EMD). The fractions were collected and concentrated using centrifugal concentrators again. The $hOGT_{4.5}$ protein was monomeric in solution, as determined by gel filtration and sedimentation equilibrium analytical ultracentrifugation. The protein was then diluted 1:1 in water before setting up crystals.

Native crystals. All crystals were grown with the hanging drop method at room temperature. For the UDP structure, 7 mg ml^{-1} protein was incubated with 1 mM UDP for several hours at 4 °C. After screening, optimal crystals were obtained when 10 μl of protein was mixed with 5 μl of reservoir solution containing 1.45 M potassium phosphate dibasic, 8 mM EDTA, and 1% xylitol. After several days, hexagonal rod crystals grew, to a maximum size of about 400 × 100 × 100 μm. Crystals were flash frozen using a cryoprotectant consisting of 1.8 M potassium phosphate and 27% xylitol. For the peptide–UDP complex, OGT was incubated with 1 mM UDP and 2 mM CKII3K peptide[28,29] for several hours at 4 °C. Crystals were obtained by mixing 8 μl protein solution with 4 μl reservoir containing 1.6 M Li_2SO_4 and 0.1 M bis-tris propane-HCl pH 7.0 (1,3-bis(tris(hydroxymethyl) methylamino)propane). Trapezoidal crystals appeared after several days. Crystals were frozen in a cryprotectant consisting of 1.72 M Li_2SO_4, 0.05 M Bis Tris Propane, pH 7.0 and 28% xylitol.

Heavy metal soaks. Several heavy metal compounds were screened using the method of ref. 31. After identifying several promising heavy metal compounds, the following conditions gave useful derivatives: K_2PtCl_4, 10 mM, 1 h soak; sodium aurothiomalate, 10 mM, 15 min soak; K_2PtCl_4, 10 mM, 10 min soak; K_2PtBr_4, 1 mM, 1 h soak.

Data collection. All the data were collected at NSLS X29 or X25 at Brookhaven National Laboratory except for the gold derivative, which was collected at ID24C at APS at Argonne National Laboratory. The heavy metal derivatives were collected at the following peak wavelengths: gold at 1.0384 Å and platinum at 1.0715 Å. The UDP structure and all the derivatives belonged to the space group *P*321. The peptide complex crystals were *I*2. All data sets were processed with iMosflm[32] and scaled using SCALA[33].

Structure determination and refinement of the OGT–UDP structure. The structure of the native OGT–UDP complex was determined by using MIRAS with the program SHARP[34]. The native data set and all the heavy atom derivative data sets were processed with iMosflm and Scala. Heavy atom sites in the K_2PtCl_4 1 h soak data set were first determined by using HKL2MAP[35]. SAD phases were then obtained with the CCP4 program Phaser[36] (Experimental Phasing). These initial phases were then used to find the heavy atom sites in the other data sets using the CCP4[37] program FFT. After obtaining all the sites, multiple isomorphous replacement with anomalous scattering (MIRAS) phases to 4.4 Å were obtained using SHARP. The figures of merit at this resolution were 0.46329 (acentric) and 0.47049 (centric). After MIRAS phasing, the map was interpretable, and we confirmed that there were four monomers in the asymmetric unit. Density modification and phase extension to 2.78 Å with NCS averaging were performed using DM, yielding a map with clear side chains. A model was built using as a guide both the structure of the bacterial homologue (using a homology model generated with Swiss Model) and the heavy atom locations. There are two loops in the intervening domain for which there is no electron density, so these residues are omitted from the model. Twelve residues are missing from one loop and four from the other. The structure was refined with CNS[38]. Initial rigid body refinement optimized the placement of the monomers and then the components of each monomer. After several iterative rounds of simulated annealing, individual *B* factor refinement, and manual adjustments using COOT[39], the UDP and waters were added, and the structure refined to an R_{work} of 21% and an R_{free} of 24%. Refinement was completed in Phenix[40] using TLS refinement[41,42], minimization, and individual *B* factor refinement to give a final R_{work} of 18.5% and R_{free} of 21.8%. Figures were prepared using Pymol[43] and CCP4mg[44].

Structure determination and refinement of the OGT–UDP–peptide complex. Data were processed with iMosflm and Scala, and the structure was determined by molecular replacement. The refined OGT–UDP structure described above was used as a search model using the Phaser molecular replacement module[45] in CCP4. Initial molecular replacement efforts showed that whereas the catalytic domain was nearly identical in the UDP and UDP–peptide cocomplexes, the orientation of the TPRs relative to the catalytic domain was noticeably different. Therefore, the model was broken into three parts: the catalytic domain and two sections of the TPR domains. Using this approach, a good map and model were obtained, which confirmed the twofold NCS present in this structure. The peptide was built by hand, as the side chains were already clear enough at this point to place the residues properly. The peptide in the crystal structure was cleaved from KKKYPGGSTPVS SANMM to YPGGSTPVSSANMM, as confirmed by mass spectrometry. The model was then refined with Phenix. As before, repeated rounds of annealing and individual *B* factor refinement were interspersed with manual adjustments in COOT. Waters were then added and sulphate ions were added after refining the waters. The structure was completed with cycles of annealing, minimization, TLS and *B* factor refinement, leading to a final structure with R_{work} of 22.4% and R_{free} of 25.2%. The crystal packing for the two complexes is described in Supplementary Fig. 9.

Kinetics. Mutants were made from the full-length ncOGT using QuickChange mutagenesis and the primers shown in Supplementary Table 6. Kinetic measurements were performed using a previously described filter binding assay[29]. Briefly, reaction mixtures containing 500 μM CKII3K peptide (KKKYPGGSTPVSSANMM), 6 μM UDP-^{14}C-GlcNAc (300 mCi $mmol^{-1}$ specific activity, American Radiochemicals), 100 nM OGT (WT or mutant protein), and buffer (125 mM NaCl, 1 mM EDTA, 20 mM potassium phosphate, pH 7.4, and 500 μM tris(hydroxypropyl)phosphine) were incubated at room temperature for 30 min. Reactions were then quenched by spotting onto the Whatman P81 phosphocellulose disks, washed three times for five minutes in 0.5% phosphoric acid, and counted by liquid scintillation counting. Reactions proceeded to $<$10% conversion under these conditions. Positive and negative controls were conducted similarly without enzyme, and positive controls were detected by liquid scintillation counting without the phosphoric acid wash step. Data were analysed based on triplicate experiments. For product inhibition experiments, substrate concentrations were used as described in Supplementary Fig. 2. Reactions were allowed to proceed for either 30 min or 60 min and performed in triplicate and analysed with linear regression using GraphPad Prism5.

Model preparation. The $hOGT_{4.5}$ construct contains the residues 313–1031 (CPTH…KPVE) of the full-length ncOGT protein. Because the first two TPR units of $hOGT_{4.5}$ overlap with the last two TPR units of the previously crystallized human TPR domain (PDB code 1W3B)[21], we superimposed each of the $hOGT_{4.5}$ structures (PDB codes 3PE3 and 3PE4) with the TPR domain to create composite models of full length human OGT. Coordinates are provided in Supplementary Data Files 1 and 2.

Molecular dynamics. The coordinates of the OGT–UDP–peptide complex were optimized in the Protein Preparation Wizard (Schrodinger 2009) where hydrogens were added; water molecules, UDP and peptide were stripped; and the structure was minimized using the OPLS2001 forcefield. The 1-μm simulation used the CHARM27 forcefield[46], and the simple point charge model for water[47]. The CHARM27 forcefield was applied to the system using the VIPARR utility. The default Desmond relaxation was performed before simulation, and molecular dynamics were run at constant temperature (300 K) and pressure (1 bar). The simulation was performed by using the program Desmond, version 2.2.9.1.0[30] compiled by SBGrid on an optimized 64-node Linux-based InfiniBand cluster and took 75 days to complete. Molecular dynamics trajectories were processed and animated with VMD[48].

Lipid (PIP) binding assays. Recombinant OGT constructs (His- or GST-tagged full-length human OGT) were overexpressed in *E. coli* and purified by affinity chromatography, using agarose beads conjugated to nickel or glutathione, respectively. PIP binding assays were performed using PIP Strips (Echelon Biosciences). Each membrane was pre-incubated for 2 h at room temperature with a blocking solution

containing 0.1% ovalbumin (for GST fusion constructs) or 3% fatty acid free BSA (for His-tagged constructs) in buffer TBST (20 mM Tris pH 8.0, 50 mM NaCl, 0.1% Tween 20). Purified OGT proteins resuspended in TBST at various concentrations (0.2–2 μM) were applied to each membrane. Washing and developing steps were performed as outlined in the manufacturer's protocols, using the same TBST described above, and protein was detected using either anti-His or anti-GST antibodies and HRP-conjugated secondary antibodies.

31. Boggon, T. J. & Shapiro, L. Screening for phasing atoms in protein crystallography. *Structure* **8,** R143–R149 (2000).
32. Leslie, A. G. W. Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4 + ESF-EAMCB Newsl. Protein Crystallogr.* **26,** 27–33 (1992).
33. Evans, P. Scaling and assessment of data quality. *Acta Crystallogr. D* **62,** 72–82 (2006).
34. de la Fortelle, E. & Bricogne, G. Maximum-likelihood heavy-atom parameter refinement for the multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol.* **276,** 472–494 (1997).
35. Pape, T. & Schneider, T. R. HKL2MAP: a graphical user interface for phasing with SHELX programs. *J. Appl. Crystallogr.* **37,** 843–844 (2004).
36. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
37. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
38. Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
39. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
40. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
41. Painter, J. & Merritt, E. A. TLSMD web server for the generation of multi-group TLS models. *J. Appl. Crystallogr.* **39,** 109–111 (2006).
42. Painter, J. & Merritt, E. A. Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr. D* **62,** 439–450 (2006).
43. DeLano, W. L. The Pymol Molecular Graphics System. (Delano Scientific, San Carlos, CA, 2002).
44. Potterton, L. *et al.* Developments in the CCP4 molecular-graphics project. *Acta Crystallogr. D* **60,** 2288–2294 (2004).
45. McCoy, A. J. Solving structures of protein complexes by molecular replacement with Phaser. *Acta Crystallogr. D* **63,** 32–41 (2007).
46. Mackerell, A. D. Jr, Feig, M. & Brooks, C. L. III. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **25,** 1400–1415 (2004).
47. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F. & Hermans, J. in *Intermolecular Forces* (ed. Pullman, B.) 331–342 (Reidel, 1981).
48. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* 14, 27–28, 33–38 (1996).

LETTER

doi:10.1038/nature09639

Exome sequencing identifies frequent mutation of the SWI/SNF complex gene *PBRM1* in renal carcinoma

Ignacio Varela[1], Patrick Tarpey[1], Keiran Raine[1], Dachuan Huang[2], Choon Kiat Ong[2], Philip Stephens[1], Helen Davies[1], David Jones[1], Meng-Lay Lin[1], Jon Teague[1], Graham Bignell[1], Adam Butler[1], Juok Cho[1], Gillian L. Dalgliesh[1], Danushka Galappaththige[1], Chris Greenman[1], Claire Hardy[1], Mingming Jia[1], Calli Latimer[1], King Wai Lau[1], John Marshall[1], Stuart McLaren[1], Andrew Menzies[1], Laura Mudie[1], Lucy Stebbings[1], David A. Largaespada[3], L. F. A.Wessels[4], Stephane Richard[5,6], Richard J. Kahnoski[7], John Anema[7], David A.Tuveson[8], Pedro A. Perez-Mancera[8], Ville Mustonen[9], Andrej Fischer[9,10], David J. Adams[11], Alistair Rust[11], Waraporn Chan-on[2], Chutima Subimerb[2], Karl Dykema[12], Kyle Furge[12], Peter J. Campbell[1], Bin Tean Teh[2,13,14], Michael R. Stratton[1,15] & P. Andrew Futreal[1]

The genetics of renal cancer is dominated by inactivation of the *VHL* tumour suppressor gene in clear cell carcinoma (ccRCC), the commonest histological subtype. A recent large-scale screen of ~3,500 genes by PCR-based exon re-sequencing identified several new cancer genes in ccRCC including *UTX* (also known as *KDM6A*)[1], *JARID1C* (also known as *KDM5C*) and *SETD2* (ref. 2). These genes encode enzymes that demethylate (*UTX*, *JARID1C*) or methylate (*SETD2*) key lysine residues of histone H3. Modification of the methylation state of these lysine residues of histone H3 regulates chromatin structure and is implicated in transcriptional control[3]. However, together these mutations are present in fewer than 15% of ccRCC, suggesting the existence of additional, currently unidentified cancer genes. Here, we have sequenced the protein coding exome in a series of primary ccRCC and report the identification of the SWI/SNF chromatin remodelling complex gene *PBRM1* (ref. 4) as a second major ccRCC cancer gene, with truncating mutations in 41% (92/227) of cases. These data further elucidate the somatic genetic architecture of ccRCC and emphasize the marked contribution of aberrant chromatin biology.

Exome sequencing based on a solution phase capture approach[5] was performed on seven cases of ccRCC, three of which carry VHL mutations, and matching normal DNAs (See Supplementary Information and Supplementary Table 1). Captured material was sequenced using 76 base pair paired-end reads on the Illumina GAIIx platform. After read alignment, variant calling was performed using a naive Bayesian classifier algorithm for substitutions and a split-read mapping approach (PinDel[6] with substantial cancer-aware output filtering) for insertion/deletions (See Supplementary Material for details). These algorithms aim to identify somatically acquired coding and splice-site variants (that is, present in the tumour but not in the matching normal), and all mutations reported here were confirmed by PCR-based capillary sequencing. In total 156 somatic mutations were identified, of which 92 were missense, 9 nonsense, 1 canonical splice site, 1 stop codon read-through, 11 frameshift and 42 synonymous (Supplementary Table 2).

In four cases truncating mutations were identified in *PBRM1*. *PBRM1* maps to chromosome 3p21 and encodes the BAF180 protein, the chromatin targeting subunit of the PBAF SWI/SNF chromatin remodelling complex[7]. The gene is comprised of six bromodomains involved in binding acetylated lysine residues on histone tails, two bromo-adjacent homology domains important in protein–protein interaction and an HMG DNA-binding domain[4]. PBAF complex-mediated chromatin remodelling is implicated in replication, transcription, DNA repair and control of cell proliferation/differentiation[4,7]. The *SMARCB1* and *BRG1* components of this complex have inactivating mutations in rhabdoid tumours[8,9] and *BRG1* mutations have been reported in several tumour types[10]. The *PBRM1* mutations included three frame-shifting insertions and a nonsense mutation; all judged to be homozygous from SNP array and mutant allele read count data. *PBRM1* was not included in our previous PCR-based sequencing screen[2] and was the only gene, apart from *VHL*, with recurrent truncating mutations in the seven cases screened. We next sequenced *PBRM1* in a further 257 RCC cases, including 36 cases of papillary, chromophobe and other non-ccRCC cancers. Truncating mutations were identified in a remarkable 88/257 (34%) (Fig. 1) of cases, all diagnosed as ccRCC (for full data see Supplementary Tables 3 and 4). *PBRM1* mutations were all found in the context of chromosome 3p loss of heterozygosity (38/38) where SNP array data was available (http://www.sanger.ac.uk/cgi-bin/genetics/CGP/cghviewer/CghHome.cgi). Two in-frame deletion mutations were identified—a predicted 6-amino-acid deletion (p.M1209_E1214delMFYKKE) in the second BAH (bromo-adjacent homology) domain likely to be involved in protein–protein interactions within the SWI/SNF complex[4] and deletion of an isoleucine codon (Ile 57) in the first bromodomain (Fig. 1). Both deletions remove amino acids conserved to *Caenorhabditis elegans* and both were in cases with 3p LOH. The ratio of nine missense to zero silent mutations suggests that a proportion of the missense mutations are likely to be pathogenic. Six of nine missense mutations occur in bromodomains and one in the second BAH domain (Fig. 1).

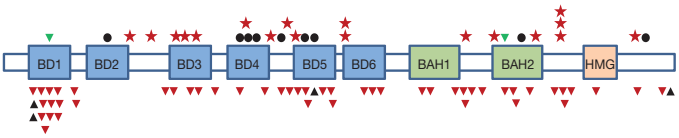


Figure 1 | PBRM1 somatic mutations. Representation of PBRM1 transcript with boxes BD1–BD6, BAH1, BAH2 and HMG indicating the positions of the bromodomains 1–6, bromo-adjacent homology domains and high-mobility group domain, respectively. Relative positions of mutations are indicated by symbols. Stars, nonsense; dots, missense; red triangles, frameshift deletions; black triangles, frameshift insertions; and green triangles, in-frame deletions. Splice-site mutations are not depicted.

[1]Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK. [2]NCCS-VARI Translational Research Laboratory, National Cancer Centre Singapore, 11 Hospital Drive, 169610, Singapore. [3]Masonic Cancer Center, University of Minnesota, Minneapolis, Minnesota 55455, USA. [4]Bioinformatics and Statistics, Department of Molecular Biology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. [5]Génétique Oncologique EPHE-INSERM U753, Faculté de Médecine Paris-Sud and Institut de Cancérologie Gustave Roussy, 94805 Villejuif, France. [6]Centre Expert National Cancer Rares INCa ''PREDIR'', Service d'Urologie, Hôpital de Bicêtre, AP-HP, 94276 Le Kremlin-Bicêtre, France. [7]Department of Urology, Spectrum Health Hospital, Grand Rapids, Michigan 49503, USA. [8]Li Ka Shing Centre, Cambridge Research Institute, Cancer Research UK, Robinson Way, Cambridge CB2 0RE, UK. [9]Bioinformatics, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK. [10]Institut für Theoretische Physik, Universität zu Köln, Zülpicherstrasse 77, 50937 Köln, Germany. [11]Experimental Cancer Genetics, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK. [12]Laboratory of Computational Biology, Van Andel Research Institute, Grand Rapids, Michigan 49503, USA. [13]Laboratory of Cancer Therapeutics, DUKE-NUS Graduate Medical School, Singapore. [14]Laboratory of Cancer Genetics, Van Andel Research Institute, Grand Rapids, Michigan, 49503, USA. [15]Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK.

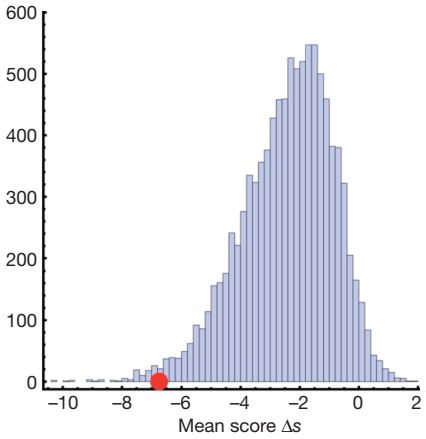


Figure 2 | Analysis of PBRM1 missense mutations. Bars represent histogram of the mean score of *in silico* generated random missense mutations (10,000 sets of three mutations that can be scored) and the red disk denotes the mean score of the somatic mutations that could be scored (T232P $\Delta s = -7.78$, A597D $\Delta s = -9.69$, H1204P $\Delta s = -2.76$). The somatic set is significantly different from the null set (*P*-value = 0.01). They have a higher negative mean score and are thus predicted to be more deleterious on average.

The bromodomains of PBRM1 have been shown to have preferential binding to different acetylated lysine configurations of histone tails, indicating they may contribute to 'reading' of the histone code[11]. The likelihood of the missense mutations having functional impact was assessed using a scoring system calibrated with protein domain alignments from Pfam (see Supplementary Methods). Three missense mutations (p.T232P, p.A597D and p.H1204P) could be scored with these alignments. This set of mutations was predicted to be deleterious, having a significantly lower mean score than a typical null set of *in silico* generated random missense mutations falling onto the scorable parts of the gene (*P*-value = 0.01 Fig. 2), making these mutations interesting candidates for functional studies.

Four *PBRM1* truncating mutations have been described in breast cancer previously[12]. Although there is frequent 3p21 LOH in small-cell lung cancer, no evidence for *PBRM1* inactivation was found[13]. To further evaluate the contribution of *PBRM1* mutation in human cancer, copy number was evaluated and the coding exons were sequenced through a series of 727 cancer cell lines of various histologies. SNP array copy number analysis (http://www.sanger.ac.uk/cgi-bin/genetics/CGP/cghviewer/CghHome.cgi) identified one homozygous deletion in the HCC-1143 breast cancer cell line, described previously[12]. Sequencing analysis identified five homozygous truncating mutations (Supplementary Table 5). Frame-shifting deletions were identified in the *VHL*-mutant A704 renal cancer, NCI-H2196 small-cell lung cancer and TGBC24TKB gall bladder cancer lines. Nonsense mutations were identified in the NCI-H226 squamous-cell lung cancer and PANC-10-05 pancreatic adenocarcinoma lines. Interestingly, a PBRM1 truncating mutation has been reported in a comprehensive pancreatic cancer mutational screen[14].

To obtain further support that *PBRM1* can act as a cancer gene, we examined data from several insertional mutagenesis screens in mice. Analyses of transposon insertion sites from a forward genetic screen performed using a conditional Sleeping Beauty transposon system[15] in a mouse pancreatic cancer model[16] revealed a significant enrichment of insertion events in *Pbrm1* amongst all genes hit using Monte Carlo simulation analyses as described previously[17]. Insertions were found in pancreatic dysplasia, intraductal (panIN) and high grade invasive tumours, indicating *Pbrm1* inactivation is an early event in this model. The mixed forward and reverse pattern of insertions is indicative of inactivation, as demonstrated by RT–PCR showing premature termination of the *Pbrm1* cDNA via splicing into the inserted transposon (Fig. 3). These data suggest that loss of *Pbrm1* cooperates with *Kras* in driving pancreatic tumour development in this model. Intriguingly, *Setd2*, previously implicated human ccRCC, was also found to rank significantly in frequency among all insertion sites and two tumours had both *Setd2* and *Pbrm1* insertions. These comparative oncogenomic data provide independent support for *PBRM1* as a cancer gene

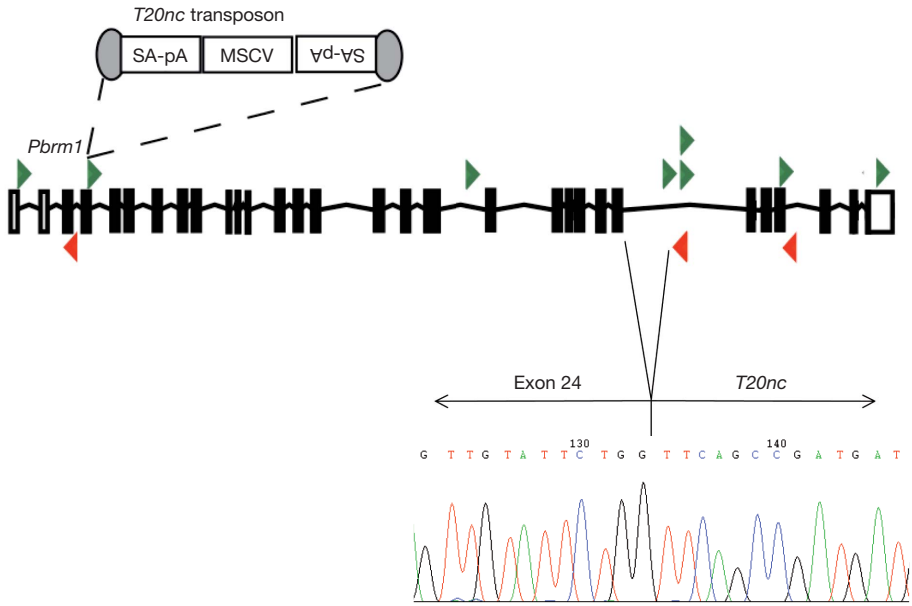


Figure 3 | *Pbrm1* is frequently mutated in a mouse model of pancreatic cancer. To identify genes that co-operate with *Kras* in the formation of pancreatic cancer a conditional allele of $Kras^{G12D}$ and *Pdx1*-Cre were combined with a conditional *Sleeping Beauty* transposase driver and the $T2Onc^{tg}$ transposon donor allele[29]. Expression of Cre results in expression of $Kras^{G12D}$ and transposon mobilization within the epithelial compartment of the pancreas. Isolation of the transposon insertion sites from a panel of 153 pancreatic cancers and pre-neoplastic lesions generated from this model revealed a common insertion site in *Pbrm1* suggesting that loss of *Pbrm1* co-operates with $Kras^{G12D}$ in pancreatic cancer development. Statistical analysis was performed as described previously[30]. Transposon insertions in the forward strand of *Pbrm1* are shown in green. Insertions in the reverse orientation are shown in red. A chromatogram from sequencing of RT–PCR products from one tumour is shown demonstrating splicing of exon 24 of Pbrm1 into the inserted transposon, thus truncating the transcript.

and suggest further investigation of the role of *PBRM1* (and *SETD2*) in human pancreatic cancer is warranted.

Abrogation of *PBRM1* expression via small interfering RNA (siRNA) knockdown in ccRCC cell lines was investigated to assess possible consequences of *PBRM1* loss. Greater than 60% knockdown of *PBRM1* RNA and protein resulted in a significant increase in proliferation 4/5 RCC lines (Fig. 4a, b and Supplementary Information). No effect was seen, however, in A704 which carries a homozygous truncating *PBRM1* mutation and expresses no PBRM1, confirming the specificity of the assay. Further, knockdown of *PBRM1* resulted in significantly increased colony formation in soft-agar and increased cell migration (Fig. 4c, d), indicative of an increase in transformed phenotype. Taken together, these data support PBRM1 having a tumour suppressor role in ccRCC.

Transcriptional profiling before and after *PBRM1* knockdown was performed using gene expression microarrays. Gene set enrichment analysis following *PBRM1* knockdown showed that PBRM1 activity regulates pathways associated with chromosomal instability and cellular proliferation (Fig. 4e and Supplementary Table 6), the latter being consistent with previous studies identifying PBRM1 as critical transcriptional regulator of p21 (also known as CDKN1A) in breast cancer cell lines[12] and work showing that PBRM1 is implicated in regulating TP53-mediated replicative senescence[18]. The PBAF complex has been shown to localize at kinetochores during mitosis[19] and SMARCB1 has been implicated in spindle checkpoint control[20], which would support the loss of PBRM1 giving rise to a chromosomal instability/spindle checkpoint expression phenotype. It may be of interest to explore further spindle checkpoint control in *PBRM1*-mutated ccRCC as a potential therapeutic opportunity.

Previous work has demonstrated that VHL loss alone is insufficient for ccRCC tumorigenesis, arguing the need for additional genetic events[21,22] (B. T. Teh, unpublished) and has further suggested the existence of a 3p21 'gatekeeper' ccRCC mutation on the basis of LOH studies[23]. The data presented here strongly suggest that inactivation of *PBRM1* comprises this second major mutation in ccRCC development. Nearly all (36/38) *PBRM1* mutant cases fall into the hypoxia signature group as described previously[2], including 13/14 cases without demonstrable *VHL* point mutations where expression data are available—further indicating the importance of *PBRM1* in typical ccRCC development. The SWI/SNF complex has been implicated in the normal cellular response to hypoxia, with impairment of the complex rendering cells resistant to hypoxia-induced cell cycle arrest[24], which would be consistent with selection for frequent loss of *PBRM1* in ccRCC. Multiple cancers have apparently concomitant *VHL*, *PBRM1* and *SETD2* mutations, with all three genes mapping to chromosome 3p, indicating that the mutations are non-redundant functionally. Half (55/107) of cases in this series with a demonstrable VHL mutation[2] have a PBRM1 mutation. Strikingly, all nine cases with a *SETD2* mutation have a mutation in either *PBRM1* or *VHL*, with 6 of 9 cases having mutations in all three genes. Physical linkage of these three ccRCC cancer genes together with their potential interaction may be the key driver for the large scale 3p LOH seen in most cases of ccRCC—being particularly parsimonious in requiring only four genetic events to unmask three tumour suppressor genes as opposed to six if the genes were on different chromosomes.

Several other mutated genes of potential interest were identified. In particular, *ARID1A* encoding the BAF250A subunit of the SWI/SNF complex was found to have two heterozygous missense mutations:

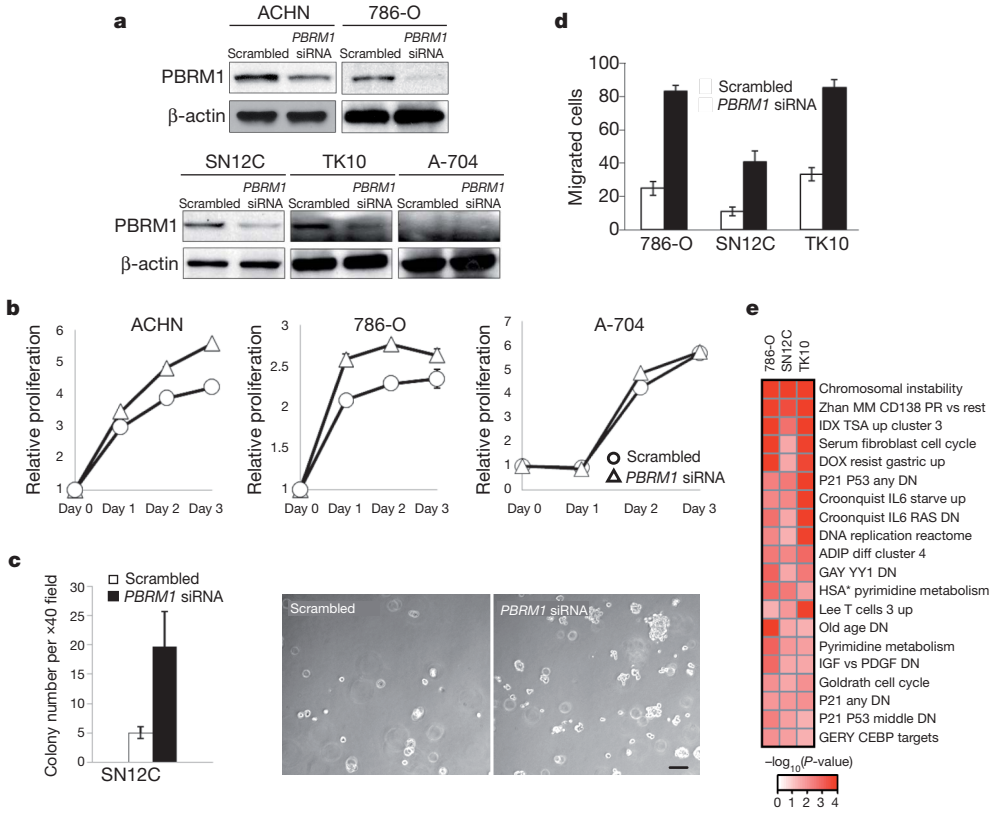


Figure 4 | Knockdown of PBRM1 expression in RCC cell lines. **a**, Verification of PBRM1 knockdown by western blotting. **b**, Silencing *PBRM1* increased the proliferation of ACHN and 786-O with wild-type *PBRM1*, but not A704 with a homozygous *PBRM1* truncating mutation. Data represent means of triplicate experiments with standard deviation, $P < 0.01$. **c**, Knockdown of *PBRM1* enhanced colony formation in SN12C cells. Data represent means of triplicate experiments with standard deviation, $P < 0.01$. Scale bar, 100 μM. **d**, Knockdown of *PBRM1* enhanced cell migration in 786-O, SN12C and TK10 cells. Data represent means of triplicate experiments with standard deviation, $P < 0.01$. **e**, Gene sets that are most significantly deregulated following *PBRM1* knockdown in three RCC cell lines using curated gene sets obtained from MSigDB (http://www.broadinstitute.org/gsea/msigdb/) and additional curated gene sets obtained from the PGSEA package (see Supplementary Information for details).

p.R1020K,c.3059G>A and p.L1872P,c.5615T>C. Both cases (PD2126, PD2127) have a PBRM1 truncating mutation. Two homozygous *ARID1A* deletions were found in SNP 6.0 data (http://www.sanger.ac.uk/cgi-bin/genetics/CGP/conan/search.cgi) in the LB1047-RCC ccRCC and NCI-SNU-5 gastric carcinoma cell lines and loss of ARID1A expression has been reported in RCC[25]. Frequent truncating ARID1A mutations have recently been reported in clear cell ovarian carcinoma[26,27]. These data all point to *ARID1A* being a cancer gene, likely to be operative in ccRCC. PD2127 was also found to have a heterozygous truncating mutation in *ARID5B*, related to *ARID1A* and recently implicated in childhood acute lymphoblastic leukaemia susceptibility[28]. The extent to which the other mutated genes identified here contribute to ccRCC will await large-scale follow-up screens. Similarly, exome and whole genome sequencing on a large number of cases is likely to yield further insights.

The identification of a second major cancer gene in ccRCC further defines the genetic and molecular architecture of this tumour type. It is remarkable that *PBRM1*, like the majority of the other non-VHL mutated cancer genes identified in ccRCC, is involved in chromatin regulation—again at least in part at the level of histone H3 modification and recognition. Understanding the contribution of *PBRM1* mutation to clinical disease progression and outcome as well the potential for exploiting SWI/SNF complex abrogation therapeutically are important future areas of renal cancer research.

METHODS SUMMARY

DNA samples from ccRCC patients tumour and matching normal were all obtained under local IRB and LREC approvals for this study and processed as described previously[2]. DNA fragmentation, library preparation and solution phase hybrid capture were according to manufacturer instructions (Agilent Technologies) and modified from protocols published previously[5]. Capillary-based Sanger sequencing for confirmations and *PBRM1* follow-up were done as described previously[2] with manual inspection of all sequencing traces. mRNA was extracted from snap-frozen mouse pancreatic lesions and subjected to RT–PCR using a nested PCR approach using primers of mouse *Pbrm1* exon 23/24 and the Carp-β-Actin Splice acceptor sequence of the T2Onc transposon cassette. Resulting bands were gel-purified and subjected to capillary-based Sanger sequencing. *PBRM1* or scrambled control siRNAs (Santa Cruz) were transfected into ccRCC cell lines using Lipofectamine 2000 (Invitrogen) according to the manufacturer's conditions. Real-time PCR and western blotting were all done using standard protocols essentially as described[1]. Expression analyses were carried out as described previously[2].

Received 28 July; accepted 2 November 2010.
Published online 19 January 2011.

1. van Haaften, G. *et al.* Somatic mutations of the histone H3K27 demethylase gene *UTX* in human cancer. *Nature Genet.* **41,** 521–523 (2009).
2. Dalgliesh, G. L. *et al.* Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature* **463,** 360–363 (2010).
3. Kouzarides, T. Chromatin modifications and their function. *Cell* **128,** 693–705 (2007).
4. Thompson, M. Polybromo-1: the chromatin targeting subunit of the PBAF complex. *Biochimie* **91,** 309–319 (2009).
5. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnol.* **27,** 182–189 (2009).
6. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25,** 2865–2871 (2009).
7. Reisman, D., Glaros, S. & Thompson, E. A. The SWI/SNF complex and cancer. *Oncogene* **28,** 1653–1668 (2009).
8. Schneppenheim, R. *et al.* Germline nonsense mutation and somatic inactivation of *SMARCA4/BRG1* in a family with rhabdoid tumor predisposition syndrome. *Am. J. Hum. Genet.* **86,** 279–284 (2010).
9. Versteege, I. *et al.* Truncating mutations of hSNF5/INI1 in aggressive paediatric cancer. *Nature* **394,** 203–206 (1998).
10. Wong, A. K. C. *et al. BRG1*, a component of the SWI-SNF complex, is mutated in multiple human tumor cell lines. *Cancer Res.* **60,** 6171–6177 (2000).
11. Chandrasekaran, R. & Thompson, M. Polybromo-1-bromodomains bind histone H3 at specific acetyl-lysine positions. *Biochem. Biophys. Res. Commun.* **355,** 661–666 (2007).
12. Xia, W. *et al.* BAF180 is a critical regulator of p21 induction and a tumor suppressor mutated in breast cancer. *Cancer Res.* **68,** 1667–1674 (2008).
13. Sekine, I. *et al.* The 3p21 candidate tumor suppressor gene BAF180 is normally expressed in human lung cancer. *Oncogene* **24,** 2735–2738 (2005).
14. Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321,** 1801–1806 (2008).
15. Keng, V. W. *et al.* A conditional transposon-based insertional mutagenesis screen for genes associated with mouse hepatocellular carcinoma. *Nature Biotechnol.* **27,** 264–274 (2009).
16. Hingorani, S. R. *et al.* Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse. *Cancer Cell* **4,** 437–450 (2003).
17. Starr, T. K. *et al.* A transposon-based genetic screen in mice identifies genes altered in colorectal cancer. *Science* **323,** 1747–1750 (2009).
18. Burrows, A. E., Smogorzewska, A. & Elledge, S. J. Polybromo-associated BRG1-associated factor components BRD7 and BAF180 are critical regulators of p53 required for induction of replicative senescence. *Proc. Natl Acad. Sci. USA* **107,** 14280–14285 (2010).
19. Xue, Y. *et al.* The human SWI/SNF-B chromatin-remodeling complex is related to yeast Rsc and localizes at kinetochores of mitotic chromosomes. *Proc. Natl Acad. Sci. USA* **97,** 13015–13020 (2000).
20. Vries, R. G. J. *et al.* Cancer-associated mutations in chromatin remodeler hSNF5 promote chromosomal instability by compromising the mitotic checkpoint. *Genes Dev.* **19,** 665–670 (2005).
21. Mandriota, S. J. *et al.* HIF activation identifies early lesions in VHL kidneys: evidence for site-specific tumor suppressor function in the nephron. *Cancer Cell* **1,** 459–468 (2002).
22. Young, A. P. *et al.* VHL loss actuates a HIF-independent senescence programme mediated by Rb and p400. *Nature Cell Biol.* **10,** 361–369 (2008).
23. Clifford, S. C., Prowse, A. H., Affara, N. A., Buys, C. H. C. M. & Maher, E. R. Inactivation of the von Hippel-Lindau (*VHL*) tumour suppressor gene and allelic losses at chromosome arm 3p in primary renal cell carcinoma: evidence for a *VHL*-independent pathway in clear cell renal tumourigenesis. *Genes Chromosom. Cancer* **22,** 200–209 (1998).
24. Kenneth, N. S., Mudie, S., van Uden, P. & Rocha, S. SWI/SNF regulates the cellular response to hypoxia. *J. Biol. Chem.* **284,** 4123–4131 (2009).
25. Wang, X. *et al.* Expression of p270 (ARID1A), a component of human SWI/SNF complexes, in human tumors. *Int. J. Cancer* **112,** 636–642 (2004).
26. Jones, S. *et al.* Frequent mutations of chromatin remodeling gene *ARID1A* in ovarian clear cell carcinoma. *Science* **330,** 228–231 (2010).
27. Wiegand, K. C. *et al. ARID1A* mutations in endometriosis-associated ovarian carcinomas. *N. Engl. J. Med.* **363,** 1532–1543 (2010).
28. Papaemmanuil, E. *et al.* Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nature Genet.* **41,** 1006–1010 (2009).
29. Collier, L. S., Carlson, C. M., Ravimohan, S., Dupuy, A. J. & Largaespada, D. A. Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature* **436,** 272–276 (2005).
30. Uren, A. G. *et al.* Large-scale mutagenesis in $p19^{ARF}$- and *p53*-deficient mice identifies cancer genes and their collaborative networks. *Cell* **133,** 727–741 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements P.A.F. and M.R.S. would like to acknowledge the Wellcome Trust for support under grant reference 077012/Z/05/Z and A. Coffey, D. Turner and L. Mamanova for assistance with the exon capture. K.F., K.D. and B.T.T. acknowledge the support of the Van Andel Research Institute. B.T.T. would like to acknowledge support from the Lee Foundation. I.V. is supported by a fellowship from The International Human Frontier Science Program Organization. D.J.A. acknowledges the support of Cancer Research UK. D.A.T. and P.A.P.-M. acknowledge the support of the University of Cambridge, Cancer Research UK and Hutchison Whampo and thank W. Howatt, A. Hazelhurst and colleagues in the CRI core facilities for their support. B.T.T. would like to dedicate this work to Tat Hock Teh.

Author Contributions I.V. and P.T. performed the main analytical aspects of the study. P.S., H.D., G.L.D., M.-L.L., G.B., C.H., L.M., S.M. performed the follow-up sequencing and analyses. K.R., D.J., J.T., A.B., C.G., D.G., M.J., C.L., J.M., A.M., L.S. contributed to the data processing, mapping and variant calling informatics. C.G. and K.W.L. performed statistical analyses. S.R., R.J.K., J.A. contributed samples and data for the clinical series. D.J.A., A.R., D.A.L., L.F.A.W., D.A.T., P.A.P.-M. performed the transposon screening and analyses. D.H., C.K.O., W.C., C.S. performed the siRNA and functional work. V.M., A.F. performed the missense mutation analysis. K.D., K.F. and J.C. performed the expression analyses. P.J.C., B.T.T., M.R.S., P.A.F. directed the study and wrote the manuscript, which all authors have approved.

Author Information Exome sequence data have been deposited at the European Genome-Phenome Archive (http://www.ebi.ac.uk/ega/) hosted by the European Bioinformatics Institute under accession EGAS00001000006 and expression data has been deposited with Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession GEO22316. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to B.T.T. (Bin.Teh@vai.org), M.R.S. (mrs@sanger.ac.uk) or P.A.F. (paf@sanger.ac.uk).

LETTER

doi:10.1038/nature09666

Interferon-γ links ultraviolet radiation to melanomagenesis in mice

M. Raza Zaidi[1], Sean Davis[2], Frances P. Noonan[3], Cari Graff-Cherry[4], Teresa S. Hawley[5], Robert L. Walker[2], Lionel Feigenbaum[4], Elaine Fuchs[6], Lyudmila Lyakh[7], Howard A. Young[7], Thomas J. Hornyak[8], Heinz Arnheiter[9], Giorgio Trinchieri[7], Paul S. Meltzer[2], Edward C. De Fabo[3] & Glenn Merlino[1]

Cutaneous malignant melanoma is a highly aggressive and frequently chemoresistant cancer, the incidence of which continues to rise. Epidemiological studies show that the major aetiological melanoma risk factor is ultraviolet (UV) solar radiation, with the highest risk associated with intermittent burning doses, especially during childhood[1,2]. We have experimentally validated these epidemiological findings using the hepatocyte growth factor/scatter factor transgenic mouse model, which develops lesions in stages highly reminiscent of human melanoma with respect to biological, genetic and aetiological criteria, but only when irradiated as neonatal pups with UVB, not UVA[3,4]. However, the mechanisms underlying UVB-initiated, neonatal-specific melanomagenesis remain largely unknown. Here we introduce a mouse model permitting fluorescence-aided melanocyte imaging and isolation following *in vivo* UV irradiation. We use expression profiling to show that activated neonatal skin melanocytes isolated following a melanomagenic UVB dose bear a distinct, persistent interferon response signature, including genes associated with immunoevasion. UVB-induced melanocyte activation, characterized by aberrant growth and migration, was abolished by antibody-mediated systemic blockade of interferon-γ (IFN-γ), but not type-I interferons. IFN-γ was produced by macrophages recruited to neonatal skin by UVB-induced ligands to the chemokine receptor Ccr2. Admixed recruited skin macrophages enhanced transplanted melanoma growth by inhibiting apoptosis; notably, IFN-γ blockade abolished macrophage-enhanced melanoma growth and survival. IFN-γ-producing macrophages were also identified in 70% of human melanomas examined. Our data reveal an unanticipated role for IFN-γ in promoting melanocytic cell survival/immunoevasion, identifying a novel candidate therapeutic target for a subset of melanoma patients.

Mechanisms associated with UV-mediated alterations to melanocytes and their microenvironment have been inscrutable because they cannot be adequately studied in cultured cells. Moreover, melanocytes represent only ~1% of skin cells, and bear few specific cell-surface markers permitting efficient isolation. To enable detailed study of melanocyte biology *in vivo*, we generated a mouse model in which expression of the reverse tetracycline-activated transactivator rtTA2s-M2, characterized by minimal leakiness and background, was regulated by the melanocyte-specific dopachrome tautomerase (*Dct*) gene promoter (Supplementary Fig. 1a). Dct-rtTA mice bred with transgenic mice bearing a histone H2B–GFP fusion construct controlled by the tetracycline response element (TRE) created Dct-rtTA/TRE-H2B–GFP bi-transgenic mice (hereafter iDct–GFP) (Supplementary Fig. 1b). iDct–GFP mice showed an inducible GFP profile from embryonic to adult stages consistent with known *Dct* expression patterns. GFP expression was observed in embryonic neural crest, retinal pigment epithelium and telencephalon, as expected (Fig. 1a and Supplementary Fig. 2). Neonatal and adult skin GFP^+ cells were strictly localized to hair follicles, where most GFP^+ cells were in bulb regions, with smaller numbers in the outer root sheath and bulge regions, harbouring melanocyte precursors[5] (Fig. 1b). Co-localization of GFP and anti-Dct antibody by immunohistochemistry unequivocally identified GFP^+ cells as melanocytes (Fig. 1c). No background GFP expression was detectable without doxycycline. Full GFP induction was achieved within 12–18 h of a single intraperitoneal injection of a non-toxic doxycycline dose in neonatal or adult mice (Supplementary Fig. 1c).

Reasoning that new clues to the molecular mechanism(s) underlying UV-induced melanomagenesis would be found within the genomic response of melanocytes to UV radiation, we used the iDct–GFP mouse to examine the responses to UVB versus UVA of melanocytes residing *in situ*, within their natural morphological and physiological microenvironment. Precisely defined wideband wavelengths and physiologically relevant doses[4] (see Methods) of UV (Supplementary Fig. 3) were used to irradiate postnatal day 1 (PD1) iDct–GFP mice, and skins were harvested at certain time points after irradiation. To avoid potential toxicity from chronic expression and interference with UV absorption, GFP was doxycycline-induced after UV irradiation, 24 h before skin harvest. GFP^+ melanocytes were phenotypically activated, characterized by raised melanocyte numbers and migration towards the epidermis as previously reported[6], 24–48 h after exposure to UVB, but not UVA, peaking at 3 days and lasting at least 10 days after irradiation (Fig. 1d and Supplementary Fig. 4a). UVB-induced melanocyte activation was specific to neonatal irradiation; adult mice irradiated at PD29 did not show this response (Supplementary Fig. 4b).

We performed an expression microarray study on melanocytes isolated from dorsal skin of iDct–GFP pups irradiated at PD1 with UVB or UVA (Fig. 1e). Doxycycline-induced GFP-labelled melanocytes were isolated via fluorescence-activated cell sorting (FACS) at 1 day and 6 days after irradiation (ages PD2 and PD7, respectively); arrays from 1 day post-UV would reflect the acute UV stress response of *in vivo* melanocytes, whereas the 6 days post-UV time point should uncover responses persisting after the acute stress response subsides. FACS isolation consistently yielded >95% melanocyte enrichment (Supplementary Fig. 5). Gene expression profiling produced robust data with good reproducibility among biological triplicates (Fig. 2a), and confirmed the absence of detectable levels of contaminating skin cell types, including keratinocytes, fibroblasts and adipocytes (Supplementary Fig. 6).

UVB elicited a potent, transient, acute stress response in melanocytes, including increased expression of p53 target genes (for example, $p21^{Waf1/Cip1}$, cyclin G1 and reprimo), whereas UVA-associated changes were subtle (Fig. 2a and Supplementary Fig. 7). Intriguingly, a small subset of genes showed a delayed response evident at 6 days after UVB exposure (Fig. 2a), including a putative IFN-responsive gene signature[7]

[1]Laboratory of Cancer Biology and Genetics, National Cancer Institute, Bethesda, Maryland 20892, USA. [2]Genetics Branch, National Cancer Institute, Bethesda, Maryland 20892, USA. [3]Laboratory of Photobiology and Photoimmunology, Department of Microbiology, Immunology and Tropical Medicine, George Washington University Medical Center, Washington, District of Columbia 20037, USA. [4]Laboratory Animal Sciences Program, National Cancer Institute, Frederick, Maryland 21702, USA. [5]Flow Cytometry Core, George Washington University Medical Center, Washington, District of Columbia 20037, USA. [6]Rockefeller University, New York, New York 10021, USA. [7]Cancer and Inflammation Program, National Cancer Institute, Frederick, Maryland 21702, USA. [8]Dermatology Branch, National Cancer Institute, Bethesda, Maryland 20892, USA. [9]National Institute of Neurological Disorders and Stroke, Bethesda, Maryland 20892, USA.

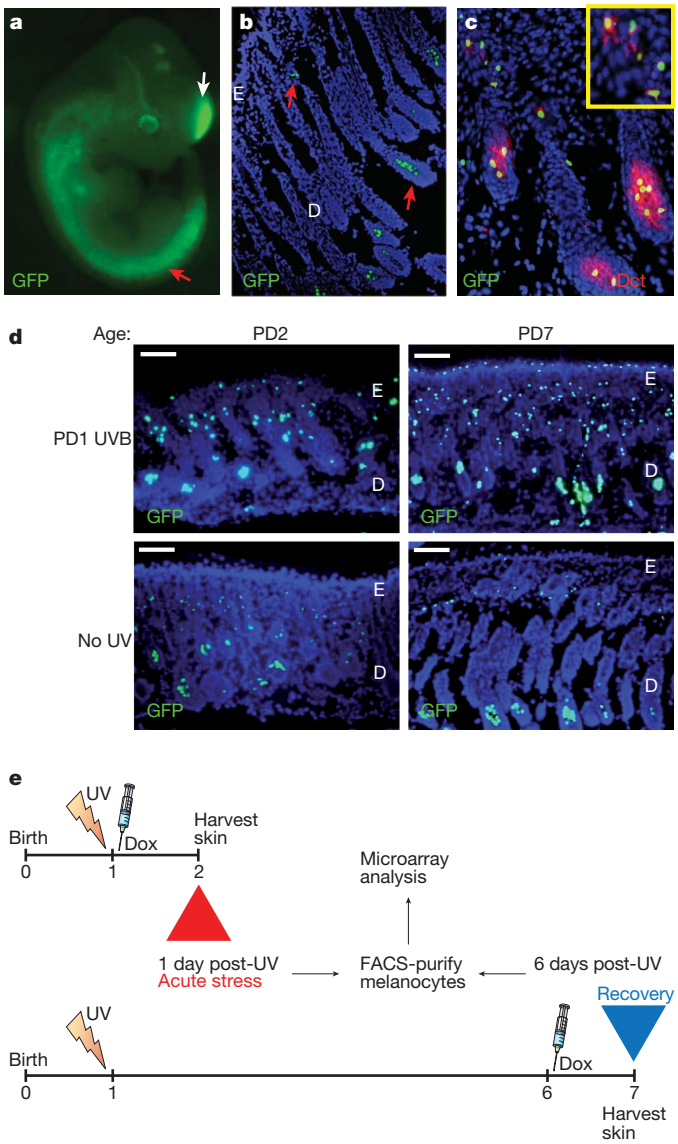


Figure 1 | Melanocyte-specific GFP expression reveals UVB-induced activation. **a**, Embryonic day 11.5 iDct–GFP embryo showing GFP$^+$ cells in neural crest (red arrow) and telencephalon (white arrow); magnification, ×10. **b**, In 7-day old pup skin GFP$^+$ cells are located in the bulb (lower arrow) and bulge (upper arrow) regions of hair follicles. Blue, 4′,6-diamidino-2-phenylindole (DAPI); magnification, ×40. D, dermis; E, epidermis. **c**, Immunohistochemistry with anti-Dct antibody shows co-localization with GFP in iDct–GFP skin; magnification, ×100. **d**, UVB-induced activation of melanocytes, characterized by proliferation and migration towards epidermis. Dorsal skins were examined at 1 day (at age PD2) and 6 days (PD7) after irradiation. Scale bars, 40 μm. **e**, Schematic of the regime for isolating GFP$^+$ melanocytes by FACS. Pups are irradiated at PD1, and dorsal skins harvested at either PD2 (24 h post-UV) or PD7 (6 days post-UV). Doxycycline (Dox) injections are always given after irradiation, 24 h before skin harvest.

(Fig. 2a and Supplementary Table 1). Four upregulated genes (*Ccl8*, *Ctla4*, *H2-K1* and *H2-T23*) from this group were validated by quantitative reverse transcription–polymerase chain reaction (qRT–PCR) (Fig. 2b). The response was neonate specific (Supplementary Fig. 8), and included genes implicated in conferring immunoevasiveness (that is, *Ctla4*, *H2-T23*, *H2-M3*, *Cfb* (also known as *H2-Bf*) and *C4a* (also known as *Slp*)).

To determine whether IFN signalling has a biologically significant role in UVB-induced melanocyte activation, we blocked both type-I (IFN-α and IFN-β) and type-II (IFN-γ) interferons by neonatal administration of anti-IFN-α receptor 1 (IFN-αR1) and anti-IFN-γ antibodies, respectively. Melanocytes from 6-day post-UVB skin were activated in the presence of isotype control antibody, whereas the anti-IFN-αR1+anti-IFN-γ antibody combination completely abolished this response (Fig. 2c). Moreover, although antibody-mediated blockade of type-I IFN-α/β signalling alone failed to overtly affect UVB-mediated activation, blockade of IFN-γ alone markedly inhibited this response (Fig. 2c). These results were corroborated by flow cytometric quantification of GFP$^+$ skin cells from each antibody-treated group (Supplementary Fig. 9). We next isolated GFP$^+$ melanocytes from UV-irradiated neonates blocked with either anti-IFN-γ or anti-IFN-αR1, and compared their expression patterns to control. Anti-IFN-γ and anti-IFN-αR1 antibodies repressed expression of a common gene set associated with the IFN response (Supplementary Fig. 10). However, IFN-γ inhibition also more potently repressed expression of several non-classical major histocompatibility complex (MHC) class Ib antigens (for example, *H2-T23* and *H2-Q2*), as well as *Psmb9*, *Gbp2*, *Icam*, *Irf1* and *Fosb*; members of this gene subset should be responsible for the observed melanocytic phenotypes. Notably, IFN-γ blockade exclusively and potently inhibited expression of chemokine Ccl8 (also known as MCP-2).

We determined the IFN-γ source by interrogating 6-day post-UVB-irradiated skin for immune-cell infiltration. Immunohistochemistry failed to detect T cells, B cells, dendritic cells, natural killer (NK), or NK-T cells; however, CD11b$^+$ cells of myeloid origin were evident (Supplementary Fig. 11). Anti-F4/80 and anti-Gr-1 antibodies identified these as macrophages (F4/80$^+$Gr-1$^-$), not neutrophils (Fig. 3a, upper panel). Nearly 90% of CD11b$^+$ cells were also F4/80$^+$ (Supplementary Fig. 12). Notably, the adult response was distinct; dorsal skins from PD35 mice 2 days after UVB showed predominant Gr-1$^+$ cell infiltration, and minimal F4/80$^+$ cells (Fig. 3a, lower panel), as reported[8].

To determine if infiltrating macrophages expressed IFN-γ, as has been suggested[9], type-I and type-II interferon mRNAs were quantified from FACS-purified CD11b$^+$ and F4/80$^+$ cells. Both had upregulated IFN-α and IFN-γ expression, and to a lesser extent IFN-β (Fig. 3b). Flow cytometry demonstrated that 28% of both CD11b$^+$ and F4/80$^+$ cells expressed IFN-γ (Fig. 3c). Our data support the notion that UVB-recruited, infiltrating macrophages secrete IFN-γ, inducing the IFN signature detected in activated melanocytes. Although we were unable to detect NK cell markers in FACS-isolated CD11b$^+$ cells by qRT–PCR (Supplementary Fig. 13), we cannot rule out a possible contribution by undetectably low numbers of NK or other inflammatory cells.

That UVB-induced chemokines were responsible for neonatal macrophage recruitment, particularly Ccr2/Ccr5 ligands, was indicated by our melanocyte microarray data and confirmed by qRT–PCR (Supplementary Fig. 14a, b). These results were further corroborated through treatment of cultured melan-c normal immortalized melanocyte cell line with non-cytocidal UV (Supplementary Fig. 14c). In contrast, upregulation of these chemokines was not detected in UV-irradiated whole skin (Supplementary Fig. 15) or isolated keratinocytes[10]. Ccr2 and Ccr5 were highly expressed in skin-infiltrating macrophages, but not in non-activated RAW264.7 macrophages (Supplementary Fig. 16). Finally, mice deficient in *Ccr2* were significantly inhibited in their ability to recruit F4/80$^+$ macrophages into neonatal skin (Fig. 3d); in contrast, *Ccr5*-deficient mice showed no significant difference (data not shown).

The arrival of IFN-γ-expressing macrophages coincided with a >100-fold upregulation in melanocyte expression of the Ccr2 ligand Ccl8, a known IFN-γ-response gene[11], whereas expression of other Ccr2 ligands had returned to baseline. We propose that recruited IFN-γ$^+$ macrophages enhance melanocyte Ccl8 expression, reinforcing macrophage–melanocyte interactions and fueling an inflammatory positive feedback loop. To confirm its ability to potently chemoattract macrophages, Ccl8 was ectopically expressed in F5061 cells—established from a UV-induced melanoma from an immunocompetent hepatocyte

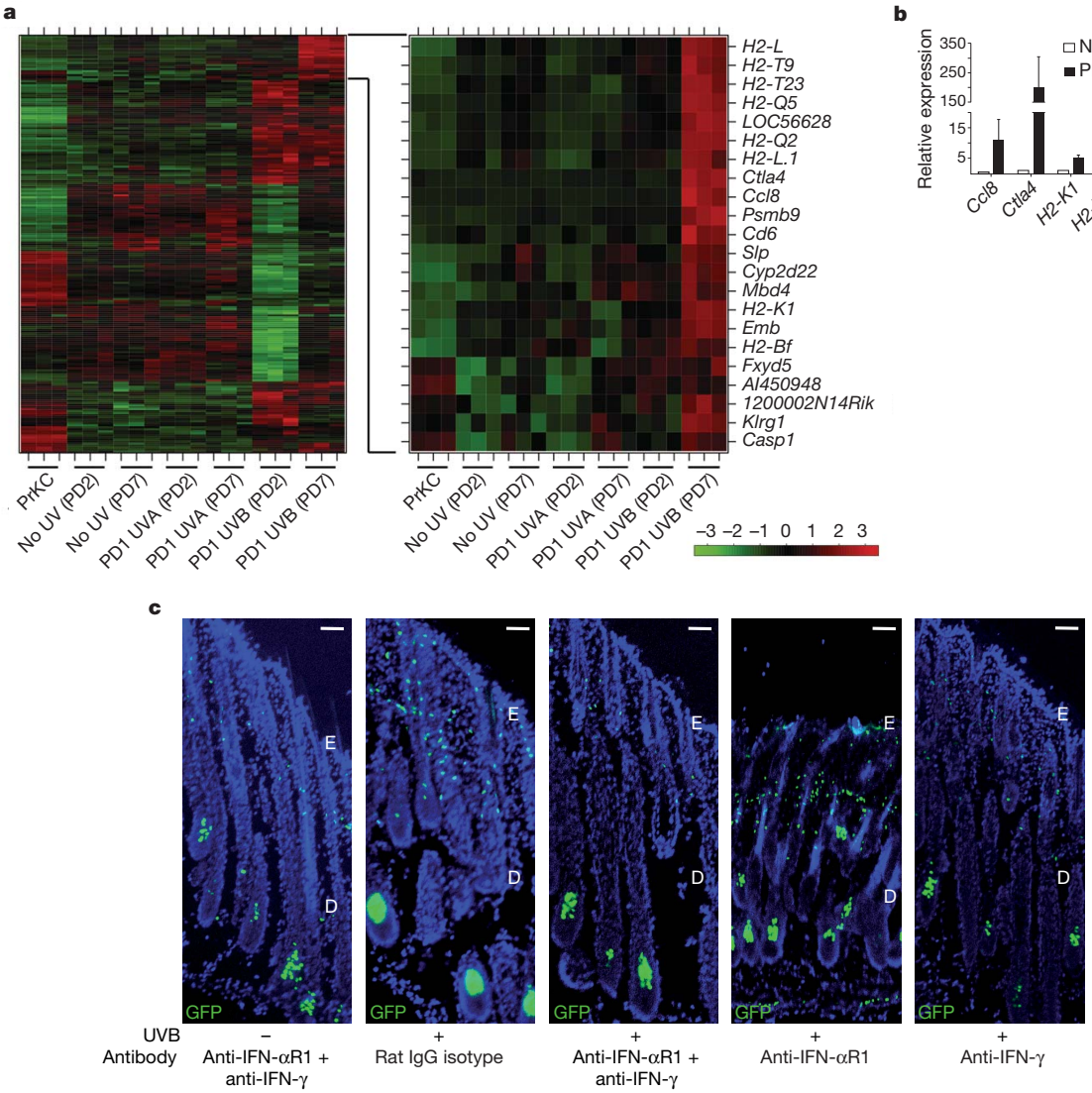


Figure 2 | UVB-induced melanocyte activation is mediated by IFN-γ. **a**, Unsupervised clustering of complementary DNA (cDNA) microarray analysis of gene expression in FACS-sorted melanocytes from 1 day (PD2) or 6 days (PD7) after UVB or UVA irradiation, and respective unirradiated controls. The expanded heatmap (right) shows the delayed induced gene subset, which includes multiple genes known to be induced by IFN-γ. Primary mouse keratinocytes (PrKC) were included as controls. All groups included biological triplicates. **b**, qRT–PCR validation of expression of 4 genes ($n = 3$ samples each) from IFN signature (error bars = s.e.m.). **c**, Antibody-mediated blockade of interferons by treating pups with intraperitoneal injections of anti-IFN-αR1, anti-IFN-γ, or both in combination, 1 h before and 3 days after UVB irradiation at PD1. The dorsal skins were harvested ($n = 3$ each group) and analysed for melanocyte activation. Representative images are shown. D, dermis; E, epidermis. Scale bars, 18 μm.

growth factor/scatter factor (HGF/SF)-transgenic mouse. F5061-Ccl8 cells were subcutaneously inoculated into syngeneic FVB/N mice, markedly enhancing macrophage infiltration into the transplantation site (Fig. 3e). Conditioned media from F5061-Ccl8 cells also significantly elevated transmembrane migration of RAW264.7 macrophages (Supplementary Fig. 17).

Macrophages show either anti- or pro-tumorigenic properties[12]. F4/80$^+$ macrophages were isolated from 6-day post-UVB neonatal skin, admixed at a ratio of 1:5 with F5061 melanoma cells and transplanted subcutaneously into FVB/N mice. Admixed transplants showed significantly increased growth relative to controls (Fig. 4a and Supplementary Fig. 18), indicating that these activated macrophages were pro-tumorigenic. In contrast, macrophages isolated from spleens of unirradiated control pups did not affect tumour growth (Supplementary Fig. 19). K_i-67 immunohistochemistry showed no difference in proliferation between admixed tumours versus controls (Supplementary Fig. 20); however, TdT-mediated dUTP nick end labelling (TUNEL) assays revealed significantly less apoptosis in admixed tumours (Fig. 4b and Supplementary Fig. 21), demonstrating that UVB-recruited macrophages promote melanoma cell survival. Immunohistochemistry confirmed that a subset of tumour-associated macrophages maintained IFN-γ expression (Supplementary Fig. 22). Moreover, macrophage presence strongly correlated with enhanced Ctla4 expression in F5061 melanoma cells (Supplementary Fig. 23), recapitulating the functional consequence of macrophage infiltration observed in neonatal skin melanocytes (Fig. 2a, b).

Despite its well-documented anti-tumorigenic activity[13], IFN-γ has also been implicated as a pro-tumorigenic factor[14,15]. To determine if macrophage-secreted IFN-γ was responsible for the enhanced melanoma growth observed in admixed tumours, we included intraperitoneal administration of either anti-IFN-γ or control antibodies. Whereas admixed melanomas in the control group showed the expected enhanced growth, those in mice given anti-IFN-γ antibody showed significantly reduced growth (Fig. 4c). Immunophenotyping of tissue microarrays containing UVB-induced mouse melanomas showed that most (66%) were macrophage-rich, with fewer having T cells (59%) and B cells (32%)

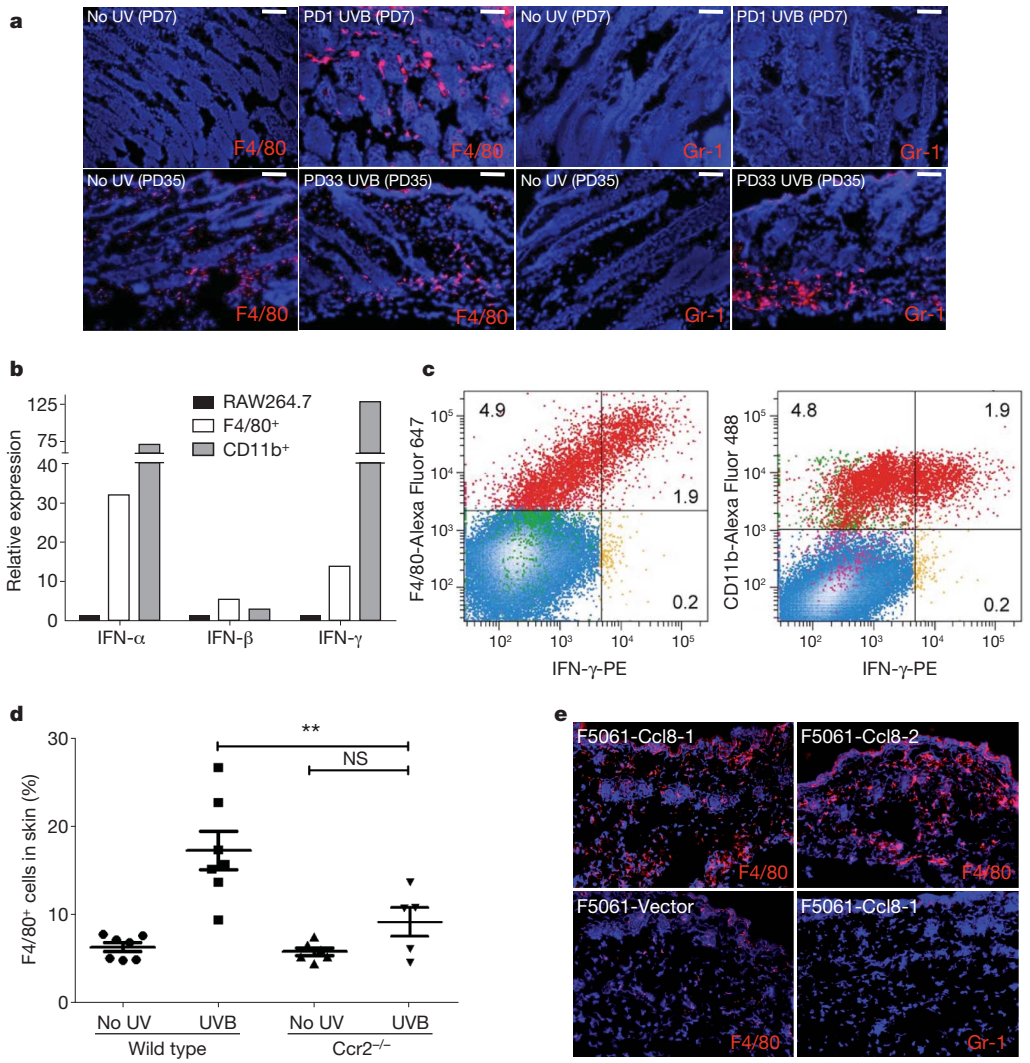


Figure 3 | UVB induces chemoattraction of IFN-γ-producing macrophages into neonatal skin. **a**, Immunohistochemistry with anti-F4/80 and anti-Gr-1 antibodies in dorsal skins of UVB-irradiated and unirradiated neonatal (upper panel) and adult (lower panel) mice. Scale bars, 40 μm. **b**, qRT–PCR for IFN-α, IFN-β and IFN-γ expression in F4/80$^+$ and CD11b$^+$ cells FACS-isolated from neonatal dorsal skins UVB-irradiated at PD1 and examined at PD7, compared with non-activated RAW264.7 macrophages. **c**, Flow cytometric analysis of PD1 UVB-irradiated (PD7 examined) skin-cell-suspension-identified IFN-γ$^+$ macrophages. PE, phycoerythrin. **d**, Flow cytometric analysis of macrophage (F4/80$^+$) infiltration into skin, 2 days after UVB irradiation in Ccr2-deficient pups (irradiated at PD1), as compared to wild-type pups. $^{**}P < 0.01$; one-way ANOVA test with post-hoc Tukey analysis. NS, not significant. **e**, F5061 melanoma cells ectopically expressing Ccl8 chemoattract F4/80$^+$ macrophages (red), but not Gr-1$^+$ cells, to sites of subcutaneous inoculation in syngeneic FVB/N mice. One vector-transfected and two Ccl8-transfected clone cells were used. Blue, DAPI.

(Supplementary Fig. 24). To determine if human melanoma-associated macrophages produce IFN-γ, we performed dual immunohistochemistry for CD68 and IFN-γ using human melanoma tissue microarrays. We discovered that 19 of 27 (70%) melanomas examined contained abundant macrophages (CD68$^+$), of which all 19 demonstrated CD68$^+$ IFN-γ$^+$ dual positivity (Fig. 4d and Supplementary Fig. 25).

In this study we show that UV incites melanomagenesis not only through DNA mutagenesis, but also by altering interactions between melanocytes and their microenvironment to regulate remodelling of UV-damaged skin. On the basis of our results, a model implicating a neonatal-specific UV-induced pro-melanomagenic inflammatory cascade emerges (Fig. 4e). In accordance with their relative tumorigenicity in albino HGF/SF-transgenic mice[4], UVB not UVA induces melanocytic expression of multiple chemoattracting Ccr2 ligands (Ccl2, Ccl7, Ccl8, Ccl12), recruiting Ccr2$^+$ macrophages into neonatal skin. IFN-γ from recruited macrophages stimulates melanocyte proliferation and migration, and the expression of genes implicated in immunoevasion/survival.

Erythemal neonatal UVB causes robust macrophage infiltration and is melanomagenic; adult skin responds with a rapid, short-lived neutrophil influx, but no melanoma. We propose that mechanisms underlying neonatal UVB-induced melanomagenesis operate within the immunoediting paradigm[13]: UVB-activated mutant neonatal melanocytes—particularly progenitors[16]—exposed to inflammation evade immune-mediated elimination, persisting through an extended equilibrium phase before evolving into clinically significant melanoma. Macrophage-induced melanocyte proliferation would more efficiently fix UV-induced mutations in prospective melanoma cells, whereas enhanced melanocyte migration could facilitate UVB-associated long-term tolerance to melanocytic antigens[17] by promoting aberrant melanocyte–immune-cell interactions. Moreover, enduring inflammation-associated epigenetic alterations occur in transformed cells[18], and perhaps in long-lived macrophage subpopulations[19], indefinitely extending their biological effects.

Notably, our systemic antibody blockade experiments demonstrating the importance of physiologically relevant IFN-γ in UVB-induced melanocyte activation and melanoma cell survival strongly support the

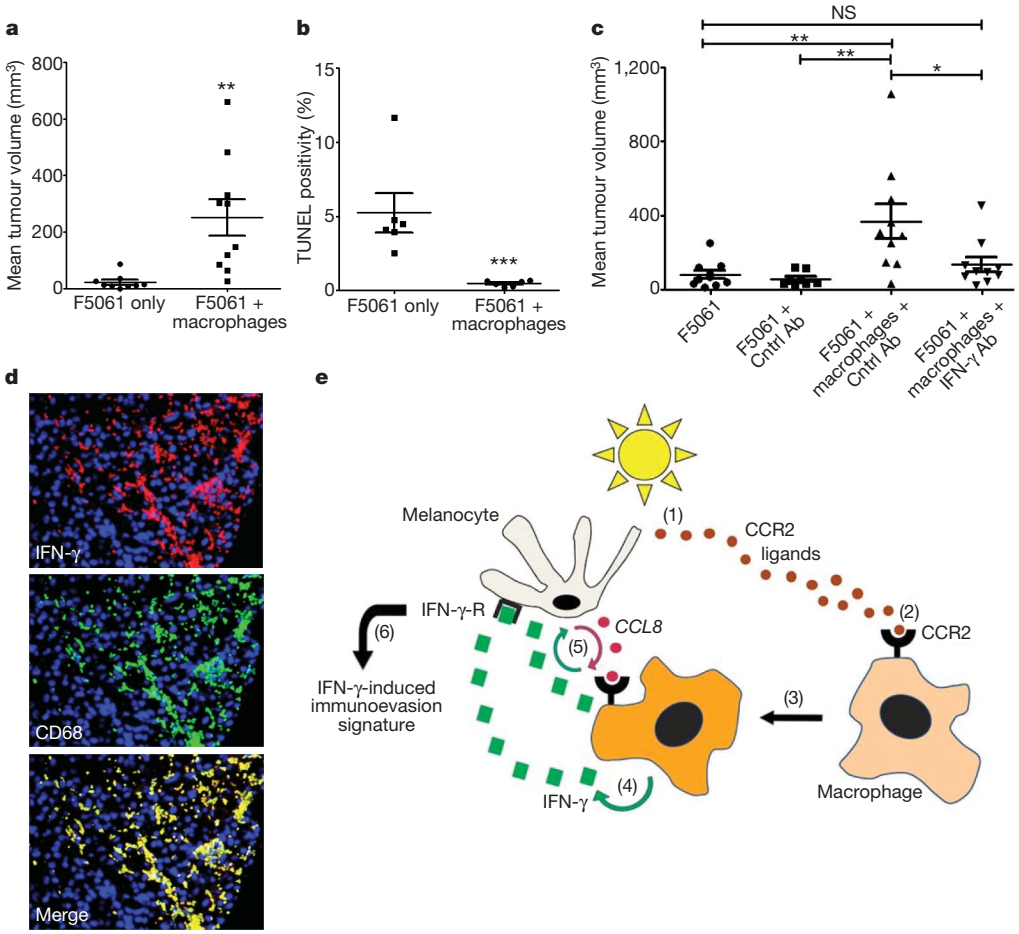


Figure 4 | IFN-γ mediates pro-tumorigenic effects of UVB-recruited skin macrophages. **a**, Mean volumes (± s.e.m.) of F5061 melanomas admixed with skin macrophages isolated from PD1 UVB-irradiated (PD7-examined) pups ($n = 10$), versus F5061-only controls ($n = 9$), following subcutaneous transplantation in syngeneic FVB/N mice. $^{**}P < 0.01$. **b**, Percentages of TUNEL plus cells (± s.e.m.) in admixed and control tumours ($n = 6$ each). $^{***}P < 0.001$. **c**, Antibody-mediated blockade of IFN-γ significantly inhibits pro-tumorigenic effects of macrophages. Cntrl Ab, control antibody. $^{*}P < 0.05$; $^{**}P < 0.01$; NS, not significant. One-Way ANOVA test with post-hoc Tukey analysis. **d**, Dual immunohistochemistry with anti-IFN-γ (red) and anti-CD68 (green) antibodies on a human melanoma tissue microarray shows IFN-γ-expressing macrophages (yellow). Representative tumour is shown. **e**, Schematic representation of the UVB-induced inflammatory cascade leading to IFN-γ-mediated immunoevasion and survival of melanocytes during sunburn-associated remodelling. UV induces release of CCR2 ligands (1); that activate CCR2^{+} macrophages (2); which are recruited to neonatal skin (3); macrophages secrete IFN-γ (4); which activates melanocytes, inducing expression of genes that include *CCL8*, fueling inflammation (5); and immunoevasion (6).

notion that IFN-γ can be pro-tumorigenic as well as anti-tumorigenic, depending on the context, intensity and durability of the IFN-γ signal[14]. In fact, an association between IFN-γ-induced genes and early mouse melanocytic lesions has been reported[15], and serum IFN-γ has been implicated as an independent prognostic indicator for melanoma recurrence[20]. We propose that IFN-γ-associated survival mechanisms operational in neonatal melanocytes are recapitulated in melanoma, contributing to the selection of more aggressive, therapeutically resistant phenotypes. Relevance to human melanoma is supported by the detection of macrophage-associated IFN-γ expression in most patient samples examined, and a clinical trial showing that IFN-γ may have adverse effects regarding melanoma patient relapse and mortality[21]. We provide the first evidence, to our knowledge, that IFN-γ-R signalling can facilitate melanoma progression, a remarkable discovery considering that high-dose IFN-α is used to treat melanoma, albeit with limited success[22]. Non-overlapping functions of type-I and type-II interferons are well described[13], and strongly supported by our data.

The IFN 'survival signature' associated with UVB-activated mouse melanocytes contains genes involved in human melanoma immunoevasion, including non-classical MHC class Ib antigens (mouse H2-M3/human HLA-G; mouse H2-T23 (also known as Qa-1)/human HLA-E)[23,24]. HLA-E suppresses NK and cytotoxic T lymphocytes[25,26]. This IFN signature also features complement isoforms C4a and C4b, implicated in systemic autoimmunity suppression[27], and CTLA4, a potent immune evasion facilitator. CTLA4 is also highly upregulated in mouse melanoma cells admixed with neonatal macrophages, and expressed on human melanoma cells, where it may be involved in immune escape[28,29].

Here we identify novel cellular/molecular inflammatory mechanisms centred on IFN-γ signalling that may underlie the initiation, survival and/or outgrowth of UVB-induced melanoma cells. We propose that such mechanisms are highly relevant to strategies used by melanoma cells to evade immunosurveillance in patients. In what could prove to be a paradigm shift, our data strongly suggest that IFN-γ/IFN-γ-R or its downstream pathway members represent promising prognostic markers and/or efficacious therapeutic targets in an appropriate subset of melanoma patients.

METHODS SUMMARY

All mouse studies were performed under the strict guidelines of Animal Study Protocols approved by the Animal Care and Use Committees at the National Cancer Institute/National Institutes of Health and the George Washington University Medical Center. The Dct-rtTA transgenic mice express rtTA2s-M2 under the control of the *Dct* (*Trp2*) gene promoter. Thirty-one founder transgenic mice were generated by standard microinjection techniques, individually crossed

with the TRE-H2B–GFP transgenic mice, and screened for a line that produced no leaky background expression of GFP without doxycycline and quickly responded to a single intraperitoneal doxycycline injection. One such line was selected to make a double homozygous Dct-rtTA/TRE-H2B–GFP transgenic mouse line in albino FVB/N background, and males were crossed with wild-type FVB/N females to produce litters that were 100% double heterozygous (iDct–GFP). One day after birth, the pups were irradiated with UV radiation (UVB, UVA, or sham)[4] (Supplementary Fig. 3). Dorsal skins were harvested at 24 h and 6 days after irradiation (or sham irradiation). Twenty-four hours before mice were killed and dorsal skins harvested, neonates were injected intraperitoneally with doxycycline at 80 $\mu g\, g^{-1}$ body weight to activate GFP expression. Skins from 6–8 pups of a litter were pooled to prepare single-cell suspensions using a published protocol[30], followed by isolation of GFP^+ melanocytes via FACS. RNA samples isolated from these melanocytes were subjected to cDNA microarray on Illumina Murine Beadchips v. 2.0 (Illumina). qRT–PCR validation was performed using primers listed in Supplementary Table 2. Immunohistochemical and flow cytometric analyses were performed using antibodies listed in Supplementary Table 3.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 26 March; accepted 9 November.
Published online 19 January 2011.

1. Garibyan, L. & Fisher, D. E. How sunlight causes melanoma. *Curr. Oncol. Rep.* **12,** 319–326 (2010).
2. Whiteman, D. C., Whiteman, C. A. & Green, A. C. Childhood sun exposure as a risk factor for melanoma: a systematic review of epidemiologic studies. *Cancer Causes Control* **12,** 69–82 (2001).
3. Noonan, F. P. *et al.* Neonatal sunburn and melanoma in mice. *Nature* **413,** 271–272 (2001).
4. De Fabo, E. C., Noonan, F. P., Fears, T. & Merlino, G. Ultraviolet B but not ultraviolet A radiation initiates melanoma. *Cancer Res.* **64,** 6372–6376 (2004).
5. Nishimura, E. K. *et al.* Dominant role of the niche in melanocyte stem-cell fate determination. *Nature* **416,** 854–860 (2002).
6. Walker, G. J. *et al.* Murine neonatal melanocytes exhibit a heightened proliferative response to ultraviolet radiation and migrate to the epidermal basal layer. *J. Invest. Dermatol.* **129,** 184–193 (2009).
7. Schroder, K., Hertzog, P. J., Ravasi, T. & Hume, D. A. Interferon-γ: an overview of signals, mechanisms and functions. *J. Leukoc. Biol.* **75,** 163–189 (2004).
8. Wolnicka-Glubisz, A. *et al.* Deficient inflammatory response to UV radiation in neonatal mice. *J. Leukoc. Biol.* **81,** 1352–1361 (2007).
9. Darwich, L. *et al.* Secretion of interferon-γ by human macrophages demonstrated at the single-cell level after costimulation with interleukin (IL)-12 plus IL-18. *Immunology* **126,** 386–393 (2009).
10. Li, D. *et al.* Rays and arrays: the transcriptional program in the response of human epidermal keratinocytes to UVB illumination. *FASEB J.* **15,** 2533–2535 (2001).
11. Proost, P., Wuyts, A. & Van Damme, J. Human monocyte chemotactic proteins-2 and -3: structural and functional comparison with MCP-1. *J. Leukoc. Biol.* **59,** 67–74 (1996).
12. DeNardo, D. G. *et al.* $CD4^+$ T cells regulate pulmonary metastasis of mammary carcinomas by enhancing protumor properties of macrophages. *Cancer Cell* **16,** 91–102 (2009).
13. Dunn, G. P., Koebel, C. M. & Schreiber, R. D. Interferons, immunity and cancer immunoediting. *Nature Rev. Immunol.* **6,** 836–848 (2006).
14. He, Y. F. *et al.* Sustained low-level expression of interferon-γ promotes tumor development: potential insights in tumor prevention and tumor immunotherapy. *Cancer Immunol. Immunother.* **54,** 891–897 (2005).
15. Aoki, H. & Moro, O. Upregulation of the IFN-γ-stimulated genes in the development of delayed pigmented spots on the dorsal skin of F1 mice of HR-1 x HR/De. *J. Invest. Dermatol.* **124,** 1053–1061 (2005).
16. Hirobe, T. Histochemical survey of the distribution of the epidermal melanoblasts and melanocytes in the mouse during fetal and postnatal periods. *Anat. Rec.* **208,** 589–594 (1984).
17. Wolnicka-Glubisz, A. & Noonan, F. P. Neonatal susceptibility to UV induced cutaneous malignant melanoma in a mouse model. *Photochem. Photobiol. Sci.* **5,** 254–260 (2006).
18. Iliopoulos, D., Jaeger, S. A., Hirsch, H. A., Bulyk, M. L. & Struhl, K. STAT3 activation of miR-21 and miR-181b-1 via PTEN and CYLD are part of the epigenetic switch linking inflammation to cancer. *Mol. Cell* **39,** 493–506 (2010).
19. Murphy, J., Summer, R., Wilson, A. A., Kotton, D. N. & Fine, A. The prolonged life-span of alveolar macrophages. *Am. J. Respir. Cell Mol. Biol.* **38,** 380–385 (2008).
20. Porter, G. A. *et al.* Significance of plasma cytokine levels in melanoma patients with histologically negative sentinel lymph nodes. *Ann. Surg. Oncol.* **8,** 116–122 (2001).
21. Meyskens, F. L. Jr *et al.* Randomized trial of adjuvant human interferon gamma versus observation in high-risk cutaneous melanoma: a Southwest Oncology Group study. *J. Natl. Cancer Inst.* **87,** 1710–1713 (1995).
22. Ascierto, P. A. & Kirkwood, J. M. Adjuvant therapy of melanoma with interferon: lessons of the past decade. *J. Transl. Med.* **6,** 62 (2008).
23. Rebmann, V., Wagner, S. & Grosse-Wilde, H. HLA-G expression in malignant melanoma. *Semin. Cancer Biol.* **17,** 422–429 (2007).
24. Derre, L. *et al.* Expression and release of HLA-E by melanoma cells and melanocytes: potential impact on the response of cytotoxic effector cells. *J. Immunol.* **177,** 3100–3107 (2006).
25. Lee, N. *et al.* HLA-E is a major ligand for the natural killer inhibitory receptor CD94/NKG2A. *Proc. Natl Acad. Sci. USA* **95,** 5199–5204 (1998).
26. Wischhusen, J., Waschbisch, A. & Wiendl, H. Immune-refractory cancers and their little helpers—an extended role for immunetolerogenic MHC molecules HLA-G and HLA-E? *Semin. Cancer Biol.* **17,** 459–468 (2007).
27. Chen, Z., Koralov, S. B. & Kelsoe, G. Complement C4 inhibits systemic autoimmunity through a mechanism independent of complement receptors CR1 and CR2. *J. Exp. Med.* **192,** 1339–1352 (2000).
28. Hodi, F. S. *et al.* Improved survival with ipilimumab in patients with metastatic melanoma. *N. Engl. J. Med.* **363,** 711–723 (2010).
29. Shah, K. V., Chien, A. J., Yee, C. & Moon, R. T. *CTLA-4* is a direct target of Wnt/β-catenin signaling and is expressed in human melanoma tumors. *J. Invest. Dermatol.* **128,** 2870–2879 (2008).
30. Wolnicka-Glubisz, A., King, W. & Noonan, F. P. $SCA\text{-}1^+$ cells with an adipocyte phenotype in neonatal mouse skin. *J. Invest. Dermatol.* **125,** 383–385 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to thank the following individuals for their support: S. Yuspa for primary keratinocytes; C. Toniatti and H. Bujard for the rtTA2sM2 construct; V. Hearing for melan-c cell line; S. Hewitt for the human melanoma tissue microarray; M. Anver for immunohistochemical staining and production/analysis of mouse melanoma tissue microarray; K. Blas and E. Vega-Valle for technical help; N. Restifo and A. Hurwitz for suggestions and discussions. This research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. M.R.Z. was supported by a National Cancer Institute Director's Innovation Career Development Award. E.C.D. and F.P.N. were supported by grants from the National Institutes of Health (awards CA53765 and CA92258), and the Melanoma Research Foundation.

Author Contributions M.R.Z. designed and performed experiments, interpreted data and wrote the manuscript. S.D. performed statistical analysis of microarray data and generated heatmaps. F.P.N. interpreted data and reviewed the manuscript. C.G.-C. managed mouse colonies. T.S.H. performed flow cytometry and FACS. R.L.W. performed cDNA microarrays. L.F. produced Dct-rtTA transgenic mice. E.F. provided TRE-H2B–GFP mice. L.L. helped design interferon blockade experiments. H.A.Y. interpreted data and reviewed the manuscript. T.J.H. evaluated GFP expression in skin and reviewed the manuscript. H.A. evaluated embryonic expression of GFP and reviewed the manuscript. G.T. designed interferon blockade experiments, provided antibodies, and reviewed the manuscript. P.S.M. designed and performed analysis of microarray data and reviewed manuscript. E.C.D. designed, measured and performed UV irradiation experiments, supervised project, and reviewed manuscript. G.M. designed experiments, interpreted data, supervised the project and wrote the manuscript.

Author Information The microarray data have been deposited in the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo) under accession number GSE25164. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to G.M. (gmerlino@helix.nih.gov) and E.C.D. (drmecd@gwumc.edu).

METHODS

Generation of Dct-rtTA and iDct–GFP mice. DNA fragments from the following plasmids were used to construct the plasmid pDct-rtTA-βGlo, which harbours the transgene fragment used to make the Dct-rtTA transgenic mice. Firstly, phβGlobin: the vector backbone is pBluescript II KS+ and the insert is a BamHI/PstI genomic fragment from the human β-globin gene, which is comprised of partial exon 2, full intron 2, full exon 3 and poly A signal. Secondly, pPDct: this plasmid contains the 3.4-kb BamHI/Eco47III fragment from the *Dct* gene promoter[31]. Thirdly, pBS/IRES-M2: this plasmid contains the rtTA2s-M2 fragment. It is a variant of pUHrT62-1[32], and was obtained from C. Toniatti and H. Bujard. The Dct-rtTA transgenic mice were generated using standard microinjection techniques in a FVB/N background strain. Dct-rtTA mice were bred with the TRE-H2B–GFP transgenic mice[33] to generate the bi-transgenic iDct–GFP mice.

Mice. Wild-type FVB/N female breeders, $Ccr2^{-/-}$ mice (strain B6.129S4-*Ccr2*^{tm1Ifc}/J), and $Ccr5^{-/-}$ mice (strain B6.129P2-*Ccr5*^{tm1Kuz}/J) were obtained from Jackson Laboratories and housed under the strict guidelines of the Institutional Animal Care and Use Committee (IACUC)-approved protocols. Transportation of mice between National Cancer Institute–Frederick and George Washington University Medical Center (GWUMC) animal facilities was performed by special animal courier service approved by IACUC.

***In vivo* UV irradiation.** iDct–GFP pups were irradiated with UV radiation (UVB, UVA, or sham) as previously described[4] (Supplementary Fig. 3). The standard erythemal dose (SED) is used to compare the sunburning effectiveness of different UV-emitting sources. These sources include UV-emitting lamps in suntanning beds, welder's arcs and sunlight among others. By determining the SED one is able to compare how efficient the UV-emitting sources are at inducing sunburn or reddening in skin relative to each other. In experimental UV studies, many laboratories use different types of UV sources for a variety of experiments. Thus, determining the SED allows one to produce equivalent amounts of 'sunburning' radiation regardless of the UV spectral output of the different sources used. The SED is produced by multiplying the spectral output or irradiance ($W\,m^{-2}$ of the UV-emitting source) with the Commission Internationale de l'Eclairage (CIE) standard erythemal action spectrum (note that $1\,W = 1\,J\,s^{-1}$). The CIE standard erythemal action spectrum is thus used to 'weight' the incoming UV radiation on an erythemal wavelength basis. The product curve produced from this convolution will give the 'erythemal effective' irradiance ($W_{\text{erythemal effective}}\,m^{-2}$) upon integration of the area under this curve. The erythemal effective irradiance is used to calculate the erythemal or sunburning dose by the equation: $\text{dose}_{\text{erythemal effective}} = \text{irradiance}_{\text{erythemal effective}} \times \text{time (s)}$. By definition, $1\,\text{SED} = 100\,J\,m^{-2}$.

Thus, our UVB source, which consists of a UVB bandpass interference filter with a half-band width of ±5 nm (dimensions: 5.1 cm × 5.1 cm) blocked to 10^{-3} to 10^{-4} outside the main bandpass (280–320 nm). This custom-made filter when coupled to our 2.5 kW Xenon arc allows for very clean wavelength resolution. The 100 cm^2 exposure area is large enough to accommodate up to 12 neonates (1–3 days old), which are situated in a three-chambered animal holder covered with a quartz lid that allows for UV transmission. Holes are drilled into the sides to allow for air exchange. This unit sits on top of a turntable rotating at approximately 3 r.p.m. to average out beam uniformity. The irradiance is measured with a spectroradiometer (StellarNet) and under standard conditions regularly produces an irradiance of approximately 0.30 CIE-effective $W\,m^{-2}$. The exposure time is constant at 90 min. These conditions regularly produce a CIE-effective dose of approximately 14–16 SED. This level of CIE-effective irradiance agrees well with our previous calculations using a radiation transfer algorithm over a global latitude/longitude grid at Northern or Southern summer mid-latitudes[34]. Direct terrestrial measurements at these latitudes indicate that under summer sunlight conditions on clear days in the Northern or Southern Hemisphere between 11:00 and 15:00 local time one can measure SEDs in the range of 10 to 20 depending on exposure conditions. As an example and as part of a project to estimate the number of SEDs one might receive during the day using, in this case, cyclists training in Spain during summer and winter, a recent paper shows just such levels in good agreement with our measurements[35].

FACS. Single-cell suspensions from pooled skin samples were prepared using a published protocol[30]. FACS was performed on a FACSAria flow cytometer (BD Biosciences) equipped with a fixed-alignment cuvette flow cell and a Coherent Sapphire solid state laser providing 13 mW of 488-nm excitation wavelength. Processing rates averaged about 5,000 events per second at 20 psi. Cells of interest were gated by a combination of forward and orthogonal light scatter, and GFP fluorescence was captured in the detector with a 530/30 bandpass filter. GFP-positive cells suspended in PBS plus 1% bovine serum albumin (BSA) were sorted as bulk populations into 12 × 75 mm polypropylene tubes. Dorsal skins from a combined litter of 6–8 pups gave a disaggregated cell suspension containing about $30–50 \times 10^6$ cells. FACS resulted in collection of $1–3 \times 10^5$ GFP-positive melanocytes per litter. Following FACS collection of cells into PBS plus 1% BSA solution, the cells were centrifuged at 4 °C, and the pelleted cells were lysed in Trizol Reagent (Invitrogen) for isolation of RNA. Total RNA was isolated from cells by Trizol organic extraction followed by RNeasy Micro Kit (Qiagen Sciences) procedure. Skin macrophages were isolated by immunostaining with anti-F4/80-Alexa Fluor 647 antibody (BioLegend)

Microarray analysis of gene expression. Three biological replicates for each group were used for cDNA microarray analyses. RNA was quantified fluorimetrically and assayed for integrity using the Agilent Bioanalyser (Agilent Technologies). One-hundred nanograms of RNA was converted to biotinylated cRNA using one round of amplification with the Illumina Labelling Kit (Illumina) and one round of T7 polymerase amplification and hybridized to Illumina Murine Beadchips v. 2.0. After hybridization and staining, the arrays were scanned in an Illumina Bead Station, and the images processed using Illumina Bead Studio software.

Statistical analysis of microarray data. The raw microarray data were extracted from the BeadStudio software and imported into the R statistical programming environment (R Development Core Team)[36] and Bioconductor[37]. A variance stabilization transformation[38] followed by quantile normalization and quality assessment were performed using the lumi package[38]. Differential expression between groups was calculated using a linear model and empirical Bayes-moderated false discovery rates (FDR) for each treated versus untreated line were calculated using the limma package[39]. A heatmap including probes showing differential expression between UVB-treated and untreated with a FDR of 0.05 was generated (Fig. 2a and Supplementary Fig. 7). Hierarchical clustering of genes was performed using single linkage and Pearson correlation distance metric. Samples were simply ordered naturally.

qRT–PCR. The most highly significant genes identified through the statistical analysis of microarray data were validated by qRT–PCR. Reverse transcription was performed on 0.1–1.0 μg RNA with Superscript III RT system (Invitrogen) following the manufacturer's protocol. Real-time PCR was performed with Quantitect SYBR Green PCR system (Qiagen Sciences) on a 7900HT Real-Time PCR machine (Applied Biosystems). 18S ribosomal RNA was used as the normalizer. The primers used in qRT–PCR are listed in Supplementary Table 2.

Histology and immunohistochemistry. Skin and tumour samples were preserved in optimal cutting temperature (OCT) compound and stored frozen at −80 °C. Five-micrometre sections were cut using a cryostat (Leica Microsystems) and were observed under a fluorescence microscope (Nikon Instruments) with DAPI (Vectashield, Vector Laboratories) counterstaining. The antibodies for fluorescence immunohistochemistry were obtained as listed in Supplementary Table 3. Haematoxylin and eosin stainings were performed using standard protocols. Select haematoxylin and eosin stained sections are shown in Supplementary Figure 26.

***In vivo* antibody-mediated blockade.** For the IFN-blockade experiment in pups, the anti-murine IFN-γ neutralizing antibody was the rat IgG clone XMG-6 ($3.19\,mg\,ml^{-1}$)[40]. The anti-murine IFN-αR1 antibody was the mouse IgG1 clone MARI-5A3 (purified monoclonal antibodies, $4\,mg\,ml^{-1}$)[41]. The control antibodies were rat IgG1 clone GL113 ($1.52\,mg\,ml^{-1}$) ascites. The *in vivo* applications of these antibodies have been described[42,43]. 0.1–0.2 mg antibody was injected per pup intraperitoneally, 1–2 h before and 3 days after UVB irradiation. For IFN-γ blockade in the tumorigenesis experiment, the anti-IFN-γ antibody was XMG-6. The control antibody was rat IgG1 against horseradish peroxidase (HRPN; BioXcell). Antibodies were administered intraperitoneally as follows: 1 mg per mouse on days −1, 0, 1, 3, 6; and 0.5 mg per mouse on days 9, 12, and 15 after inoculation. Tumours were harvested at day 18.

Immunofluorescence flow cytometry and FACS. The immunofluorescence-based flow cytometry and FACS sorting were done using antibodies listed in Supplementary Table 3, using standard procedures.

***In vitro* UV irradiation.** The melan-c melanocyte cell line (gift from V. Hearing) was irradiated with the Junggust Box, a hand-made enclosed module that houses two F20 sunlamps outputting 60% UVB and 38% UVA wavebands. A total UV dose of $175\,J\,m^{-2}$ was given at a dose rate of $2.0\,W\,m^{-2}$. The cells were irradiated in 10-cm culture dishes at 80% confluence overlaid with 3 ml PBS and the dish covered with plastic saran wrap to filter out any UVC. Following irradiation, the cells were incubated for the desired time in normal media (RPMI1640, 10% FBS, $100{,}000\,U\,l^{-1}$ penicillin, $100\,mg\,l^{-1}$ streptomycin sulphate, 2 mM glutamine and 200 nM 12-*O*-tetradecanoylphorbol-13-acetate (TPA)).

Chemotactic assays for Ccl8. The cDNA for mouse Ccl8 was amplified by RT–PCR and cloned into the pcDNA3.1 expression vector (Clontech). F5061 cells were transfected using FuGene HD (Roche). Empty pcDNA3.1 vector was transfected as control. *In vitro* chemotactic assay was performed on the RAW264.7 macrophage cell line in a Transwell system with 8-μm pore size (Corning). 1×10^5 cells were seeded in the top well, and the migrant cells were counted following haematoxylin staining. The experiment was done in duplicate. Five random microscopic fields from each replicate were counted for the number of migrant cells. For the *in vivo* chemotactic assay, 2.5×10^5 F5061-Ccl8 or F5061-vector cells were inoculated

subcutaneously in syngeneic FVB/N mice. Skin samples from the area of the inoculation were harvested two days after inoculation, preserved in OCT compound, cryosectioned, and immunohistochemistry with anti-F4/80 antibody was performed to assess macrophage infiltration.

Tumorigenicity assay. F4/80$^+$ macrophages were FACS-isolated from dorsal skins of pups 6 days after UVB irradiation at PD1. Macrophages were admixed with F5061 melanoma cells at a 1:5 ratio (0.5×10^5 macrophages: 2.5×10^5 F5061) and inoculated subcutaneously in syngeneic FVB/N mice ($n = 10$). Control mice were inoculated with 2.5×10^5 F5061 cells without macrophages ($n = 10$). Tumours were harvested after three weeks and were measured in three dimensions. Tumour volumes were calculated by the formula: V = length × width × depth × $\pi \times 1/6$. For the tumorigenicity experiment with IFN-γ blockade, for each of the two groups with macrophages, four mice had 0.5×10^5 macrophages but another 6 mice from each group had 0.42×10^5 macrophages mixed with 2.5×10^5 F5061 cells. No differences in tumour growth were detected between the mice with these differing numbers of macrophages. All four groups contained ten mice, but two mice from the F5061 plus control-antibody group died prematurely and were taken out of the analysis.

TUNEL assay. The TUNEL assay was performed using TACS TdT Fluorescein kit (R&D Systems) following the manufacturer's protocol.

Human and mouse melanoma tissue microarrays. The human melanoma tissue microarray was obtained from S. Hewitt at the National Cancer Institute/National Institutes of Health. Antigen retrieval was performed by microwave boiling in citrate buffer (pH 6). Immunohistochemistry with anti-CD68 and anti-IFN-γ antibodies (Supplementary Table 3) was performed serially with appropriate fluorescence-tagged secondary antibodies. The mouse melanoma tissue microarray has been generated by our laboratory and contains melanoma tissues from our HGF/SF transgenic mouse model, induced by a single neonatal UV radiation dose. Anti-F4/80 (macrophages), anti-CD3 (pan T cells) and anti-B220 (B cells) antibodies were used in conjunction with the Vectastain Elite ABC system (Vector Laboratories) following the manufacturer's protocol.

Statistical analyses. All statistical analyses were performed using Graphpad Prism software. Two-tailed Student's *t*-test and ANOVA with post-hoc Tukey analyses were performed as indicated to validate significant differences. Means ± s.e.m. are indicated for all statistical analyses. A P value of <0.05 was considered statistically significant.

31. Budd, P. S. & Jackson, I. J. Structure of the mouse tyrosinase-related protein-2/dopachrome tautomerase (*Tyrp2/Dct*) gene and sequence of two novel slaty alleles. *Genomics* **29,** 35–43 (1995).
32. Urlinger, S. *et al.* Exploring the sequence space for tetracycline-dependent transcriptional activators: novel mutations yield expanded range and sensitivity. *Proc. Natl Acad. Sci. USA* **97,** 7963–7968 (2000).
33. Tumbar, T. *et al.* Defining the epithelial stem cell niche in skin. *Science* **303,** 359–363 (2004).
34. De Fabo, E. C., Noonan, F. P. & Frederick, J. E. Biologically effective doses of sunlight for immune suppression at various latitudes and their relationship to changes in stratospheric ozone. *Photochem. Photobiol.* **52,** 811–817 (1990).
35. Serrano, M. A., Canada, J. & Moreno, J. C. Erythemal ultraviolet exposure of cyclists in Valencia, Spain. *Photochem. Photobiol.* **86,** 716–721 (2010).
36. Team, R. D. C. *R: A Language and Environment for Statistical Computing.* (R Foundation for Statistical Computing, 2008).
37. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).
38. Du, P., Kibbe, W. A. & Lin, S. M. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24,** 1547–1548 (2008).
39. Gentleman, R. *Bioinformatics and Computational Biology Solutions using R and Bioconductor* (Springer, 2005).
40. Cherwinski, H. M., Schumacher, J. H., Brown, K. D. & Mosmann, T. R. Two types of mouse helper T cell clone. III. Further differences in lymphokine synthesis between Th1 and Th2 clones revealed by RNA hybridization, functionally monospecific bioassays, and monoclonal antibodies. *J. Exp. Med.* **166,** 1229–1244 (1987).
41. Sheehan, K. C. *et al.* Blocking monoclonal antibodies specific for mouse IFN-alpha/beta receptor subunit 1 (IFNAR-1) from mice immunized by *in vivo* hydrodynamic transfection. *J. Interferon Cytokine Res.* **26,** 804–819 (2006).
42. Goldszmid, R. S. *et al.* TAP-1 indirectly regulates CD4$^+$ T cell priming in *Toxoplasma gondii* infection by controlling NK cell IFN-γ production. *J. Exp. Med.* **204,** 2591–2602 (2007).
43. Gramzinski, R. A. *et al.* Interleukin-12- and gamma interferon-dependent protection against malaria conferred by CpG oligodeoxynucleotide in mice. *Infect. Immun.* **69,** 1643–1649 (2001).

LETTER

doi:10.1038/nature09696

Entanglement in a solid-state spin ensemble

Stephanie Simmons[1], Richard M. Brown[1], Helge Riemann[2], Nikolai V. Abrosimov[2], Peter Becker[3], Hans-Joachim Pohl[4], Mike L. W. Thewalt[5], Kohei M. Itoh[6] & John J. L. Morton[1,7]

Entanglement is the quintessential quantum phenomenon. It is a necessary ingredient in most emerging quantum technologies, including quantum repeaters[1], quantum information processing[2] and the strongest forms of quantum cryptography[3]. Spin ensembles, such as those used in liquid-state nuclear magnetic resonance[4,5], have been important for the development of quantum control methods. However, these demonstrations contain no entanglement and ultimately constitute classical simulations of quantum algorithms. Here we report the on-demand generation of entanglement between an ensemble of electron and nuclear spins in isotopically engineered, phosphorus-doped silicon. We combined high-field (3.4 T), low-temperature (2.9 K) electron spin resonance with hyperpolarization of the ^{31}P nuclear spin to obtain an initial state of sufficient purity to create a non-classical, inseparable state. The state was verified using density matrix tomography based on geometric phase gates, and had a fidelity of 98% relative to the ideal state at this field and temperature. The entanglement operation was performed simultaneously, with high fidelity, on 10^{10} spin pairs; this fulfils one of the essential requirements for a silicon-based quantum information processor.

Most quantum information processing algorithms applied to spin ensembles have been implemented in a regime of weak spin polarization. However, owing to the very low purity of the states used, any exponential enhancement offered by quantum mechanics disappears when the scaling of total resources is considered. Highly mixed, or weakly initialized, ensembles are often interpreted as the sum of a perfectly mixed component (given by a normalized identity matrix in the density matrix representation) and a small amount, ε, of a pure component, ρ_0; thus, $\rho_{\mathrm{true}} = (1-\varepsilon)\hat{\mathrm{I}}/d + \varepsilon\rho_0$, where d is the dimensionality of the state. The $\hat{\mathrm{I}}$ component is invariant under unitary operations and is not directly observable by magnetic resonance, which produces measurements of the population differences across allowed electron and nuclear spin transitions. It is therefore straightforward to ignore the maximally mixed component: this approach is called the 'pseudo-pure approximation'[6].

There are a number of entanglement witnesses or monotones that can distinguish entangled states from (classical) separable ones. A widely used test is the positive partial transpose (PPT) criterion, which is both a necessary and a sufficient test of entanglement for two coupled, spin-1/2 particles[7,8]. Applying this test to the mixed state above, ρ_{true}, it can be shown[8] that the minimum value of ε which permits the overall state to be entangled is 1/3.

Typical values for ε in liquid-state nuclear magnetic resonance and electron spin resonance (using 10-GHz excitation at a temperature of 5 K) are $\sim 10^{-5}$ and $\sim 10^{-2}$, respectively. These values are well below the required threshold for the PPT test. Thus, although experiments performed in this regime provide a valuable test bed for techniques in entanglement generation and detection[9], the states created are only pseudo-entangled, and are fully separable. (A notable exception was the use of chemical methods to generate highly polarized hydrogen spin pairs[10], though that is a single-shot experiment with limited scalability.) To overcome this limit, we require states of higher initial purity and a method to measure the $\hat{\mathrm{I}}$ component of the density matrix.

We follow a hybrid approach, using both the electron spin and the nuclear spin associated with a phosphorus donor in silicon. Isolated donors in isotopically engineered semiconductors are of particular interest as they possess excellent decoherence characteristics (both the electron and the nuclear coherence times, T_2, exceed seconds[11,12]), can be controlled with high fidelity using microwave and radio-frequency pulses[13,14], and are promising for integrating quantum technologies into conventional semiconductor devices[15].

Neglecting the weak polarization of the nuclear spin, the initial state populations are determined by the electron spin Zeeman energy, as shown in Fig. 1a, where $\alpha = \exp(-g\mu_{\mathrm{B}}B/k_{\mathrm{B}}T)$, g is the electron g-factor, μ_{B} is the Bohr magneton, k_{B} is Boltzmann's constant, and B and T are the experimental magnetic field and temperature, respectively. At a high magnetic field (3.4 T) and low temperature (2.9 K), the donor electron spin is thermally polarized to ~66%; however, the ^{31}P nuclear spin, with a much weaker magnetic moment, has only ~0.04% polarization. Various methods, collectively known as dynamic nuclear

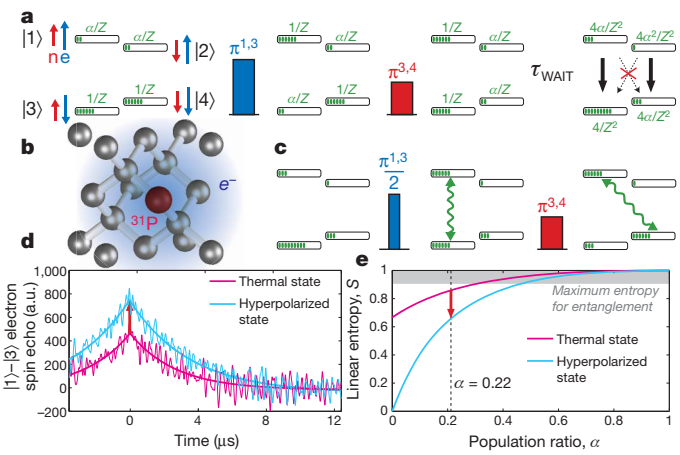


Figure 1 | Sequences for nuclear spin hyperpolarization and entanglement generation for this coupled $S = 1/2$, $I = 1/2$ spin system. a, The initial state is at thermal equilibrium, where populations (green) are distributed according to the electron spin (e) polarization at this magnetic field and temperature (see text). A pair of applied microwave and radio-frequency π pulses move spin populations to favour the $|\uparrow\rangle$ nuclear spin (n) state. After some time, $\tau_{\mathrm{WAIT}} \gg T_{1\mathrm{e}}$, there is a significant majority population in state $|3\rangle$, or $|\uparrow\downarrow\rangle$ (where the first and second arrows indicate the nuclear and electron spins, respectively). Nuclear spin and cross-relaxation processes occur on timescales much longer than $T_{1\mathrm{e}}$. **b**, Illustration of the ^{28}Si:P coupled spin system. **c**, Starting from the hyperpolarized state in **a**, an electron spin coherence is generated and transformed into the final entangled state, containing a superposition of $|\uparrow\uparrow\rangle$ and $|\downarrow\downarrow\rangle$. **d**, The growth in the electron spin echo intensity measured on the $|1\rangle$–$|3\rangle$ transition provides a measure of the population ratio, α. a.u., arbitrary units. **e**, This hyperpolarization sequence minimizes the linear entropy of the two-spin state for a given value of α.

[1]Department of Materials, Oxford University, Oxford OX1 3PH, UK. [2]Leibniz-Institut für Kristallzüchtung, 12489 Berlin, Germany. [3]PTB Braunschweig, 38116 Braunschweig, Germany. [4]VITCON Projectconsult GmbH, 07743 Jena, Germany. [5]Department of Physics, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada. [6]School of Fundamental Science and Technology, Keio University, Yokohama, 3-14-1 Hiyoshi, 223-8522, Japan. [7]CAESR, Clarendon Laboratory, Oxford University, Oxford OX1 3PU, UK.

polarization[16,17], exist for indirectly transferring electron spin polarization to the nuclear spin and often exploit cross-relaxation processes involving simultaneous electron and nuclear spin flips. Here we exploit the relative absence of cross-relaxation leading to a substantial difference in the relaxation times of the electron and nuclear spins[13], to hyperpolarize the nuclear spin rapidly and with high efficiency. This hyperpolarization process is similar to 'algorithmic cooling' methods, whereby a particular quantum bit (qubit) relaxes quickly owing to coupling to a heat bath[18].

Figure 1 illustrates our method for tackling the twin challenges of measuring and minimizing the $\hat{\mathrm{I}}$ component in the density matrix of the coupled electron–nuclear spin system. The hyperpolarization of the nuclear spin can be understood as a SWAP operation (which interchanges the states of two qubits) with the (thermally polarized) electron spin, using a combination of resonant microwave and radio-frequency π pulses. This is followed by a delay τ_{WAIT}, which is substantially longer than the electron spin relaxation time, $T_{1\mathrm{e}}$ (specifically, $\tau_{\mathrm{WAIT}} \approx 8T_{1\mathrm{e}}$), during which the electron spin relaxes back to thermal equilibrium. On this timescale, other relaxation processes (such as pure nuclear spin flips or electron–nuclear spin flip-flops) are orders of magnitude slower and can be neglected. The resulting hyperpolarized state is

$$\rho = \frac{4}{Z^2}\left(\alpha|1\rangle\langle 1| + \alpha^2|2\rangle\langle 2| + |3\rangle\langle 3| + \alpha|4\rangle\langle 4|\right)$$

where $Z = 2(1+\alpha)$ is a normalizing constant.

Although spin echo sequences can only be used to probe the population differences across energy levels, we can obtain a direct measure of the population ratio, α, by measuring the electron spin echo amplitude between levels $|1\rangle$ and $|3\rangle$ before and after the hyperpolarization sequence, as shown in Fig. 1d. Owing to the enhanced polarization of the nuclear spin, a spin echo measured on this transition increases by a factor of $2/(\alpha+1)$ in comparison with the measurement from a fully relaxed thermal state. This measure is strictly conservative: it places a lower bound on the true polarization of the electron, as imperfections such as pulse errors or residual relaxation processes only lead to a lower apparent state purity. Using this measure, we observe an enhancement of the echo intensity by a factor of 1.643(2), corresponding to an upper bound of $\alpha \leq 0.217(2)$.

Linear spin entropy (defined as $\mathcal{N}[1-\mathrm{Tr}(\rho^2)]/(\mathcal{N}-1)$ for an $\mathcal{N}$-dimensional Hilbert space) is a useful characterization of a state's purity, and ranges from one, for maximally mixed states, to zero, for pure states. Our hyperpolarization sequence corresponds to a decrease in linear spin entropy, made possible by the open quantum system's contact with the lattice heat bath (Fig. 1e). Importantly, this approach leads to the minimum possible linear entropy given the electron spin polarization resource and type of relaxation present[18]. Entanglement is maximized in a mixed, two-qubit density matrix by first minimizing the linear entropy and then generating an entangled coherence across the levels with the largest and second-smallest populations[19,20]. Following this strategy, we create an entangled state using a coherence-generating microwave $\pi^{1,3}/2$ pulse (where the superscript denotes the pair of levels addressed by the pulse) followed by a radio-frequency $\pi^{3,4}$ pulse (Fig. 1c), yielding the target state:

$$\rho = \frac{1}{2Z^2}\begin{pmatrix} 1+\alpha & 0 & 0 & 1-\alpha \\ 0 & 2\alpha^2 & 0 & 0 \\ 0 & 0 & 2\alpha & 0 \\ 1-\alpha & 0 & 0 & 1+\alpha \end{pmatrix}$$

This density matrix is entangled according to the PPT criterion when $\alpha \leq 0.432$; other preparation methods (such as pseudo-pure state preparation) require substantially higher polarization (Supplementary Information).

Having prepared the initial state and performed an entangling operation, we now use density matrix tomography to extract the final two-spin state. Owing to the weak magnetic moment of nuclear spins and necessarily low donor concentration in our sample, we are restricted to non-projective measurements of the electron spin ensemble along the σ_x and σ_y Pauli bases, which can be performed selectively on the m_I state of the nuclear spin (in product operator formalism, these bases can be written as $S_{x,y}I^{\alpha,\beta}$).

Diagonal elements of the density matrix (corresponding to state populations) are obtained by mapping pairs of population differences into an electron spin echo on the $|1\rangle$–$|3\rangle$ transition ($S_{x,y}I^{\alpha}$). The accurate detection of off-diagonal elements (coherences) is a more elaborate process, made by selectively labelling the coherence between each pair of eigenstates with a distinguishable, time-varying phase[9]. By this process, a particular phase accumulation rate provides the signature of a particular coherence, allowing the off-diagonal elements to be reconstructed from the amplitudes in the Fourier transform of a measured signal.

Here we follow an approach inspired by the Aharonov–Anandan geometric phase gate[21,22] to apply arbitrary phases in a fixed time to the four different eigenstates, and thus separately label each of the possible coherences. We apply two π pulses, along different axes, across a transition between a pair of eigenstates. The phase acquired by each eigenstate is opposite and equal to half the solid angle of its trajectory on the Bloch sphere (Fig. 2a). Thus, applying $\pi_0^{1,3}$ followed by $-\pi_\phi^{1,3}$ (subscripts denote pulse phase and, thus, nominal rotation axis) leads eigenstates $|1\rangle$ and $|3\rangle$ to assume trajectories of equal and opposite solid angle, $\pm 2\phi$. A similar operation, $\pi_0^{3,4}$ followed by $-\pi_\sigma^{3,4}$, is applied to the nuclear spin transition, such that the total operator describing the action of these four pulses is

$$U(\phi,\sigma) = \begin{pmatrix} \mathrm{e}^{-i\phi} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \mathrm{e}^{i(\sigma+\phi)} & 0 \\ 0 & 0 & 0 & \mathrm{e}^{-i\sigma} \end{pmatrix}$$

The value of ϕ is incremented by $\delta\phi$ on each shot of the experiment, with effective frequency $\nu_\phi = 2\pi/\delta\phi$ (and similarly for σ, $\delta\sigma$ and ν_σ). We then map each off-diagonal element of the density matrix in turn into $S_{x,y}I^{\alpha}$ using a set of appropriate microwave and radio-frequency π

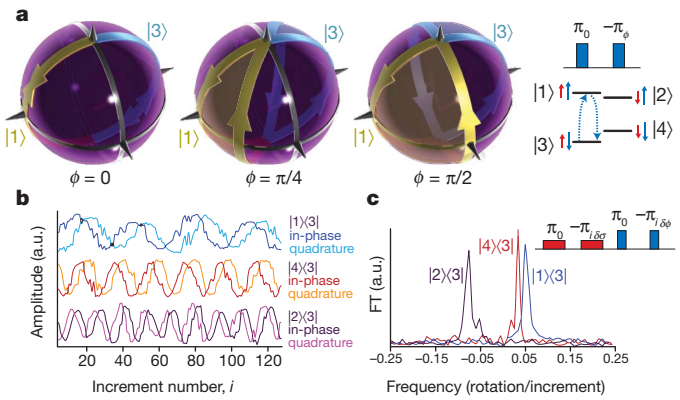


Figure 2 | Electron and nuclear spin phase rotations reveal the off-diagonal elements of the density matrix. **a**, Under the application of two consecutive $\pi^{1,3}$ pulses around different axes (ϕ), the eigenstates $|1\rangle$ and $|3\rangle$ undergo closed trajectories on the Bloch sphere with equal and opposite solid angles, $\Omega = \pm 2\phi$. Each state picks up a phase equal to half this solid angle. **b**, This π_0, $-\pi_\phi$ phase gate is applied to both electron $|1\rangle$–$|3\rangle$ and nuclear $|3\rangle$–$|4\rangle$ transitions, where the two phases are varied by different increments, $\delta\phi$ and $\delta\sigma$, as the experiment is repeated. Example oscillations are shown for three experiments where we generate an electron coherence, $|1\rangle\langle 3|$, a nuclear coherence, $|4\rangle\langle 3|$ and a zero quantum coherence, $|2\rangle\langle 3|$. **c**, Fourier transforms (FT) of the oscillations with respect to increment number show peaks located at the frequencies 0.050(8), 0.031(5) and −0.079(8), in agreement with the frequencies that were set, $\nu_\phi = 2\pi/\delta\phi = 0.05$ and $\nu_\sigma = 2\pi/\delta\sigma = 0.03$.

pulses, and measure the amplitude of the Fourier component at the effective frequency corresponding to that coherence. Quadrature measurement allows us to discriminate between positive and negative frequencies. The presence of other Fourier peaks would be illustrative of pulse errors in the mapping sequence, but as seen in Fig. 2b, c, such errors are negligible even in the absence of operations such as phase cycling.

By combining our measurements of the identity component and the diagonal and off-diagonal elements of the density matrix of the electron-nuclear spin system, we obtain the following expression for ρ:

$$\begin{pmatrix} 0.382 & 0.003+0.000i & -0.035-0.039i & 0.272 \\ 0.003-0.000i & 0.017 & -0.000+0.001i & 0.001+0.003i \\ -0.035+0.039i & -0.000-0.001i & 0.174 & -0.055-0.042i \\ 0.272 & 0.001-0.003i & -0.055+0.042i & 0.427 \end{pmatrix}$$

This state has a minimum eigenvalue under the PPT test of $-0.19(1)$ and a concurrence, C, of 0.43(4), each of which confirms the presence of finite entanglement. The results of this tomography process are shown in Fig. 3. The fidelity of the measured density matrix with respect to the target state, given that $\alpha = 0.217$, is 98.2(2)%, and is 68(2)% with respect to an ideal Bell state ($\alpha = 0$). To obtain the uncertainty in these values, we used Monte Carlo generation of physical density matrices based on the standard error of each matrix element due to noise (Supplementary Information).

The finite entanglement shown can offer direct advantages over classical methods in applications such as quantum sensors[23]. To achieve higher-purity entangled states, we could use lower temperatures; for example, we would expect $C \approx 0.99$ if these experiments were performed at 0.8 K. Complementary to this approach, entanglement purification could be performed using a larger Hilbert space at each node[24], for example using a donor atom with a higher nuclear spin (such as bismuth, with $I = 9/2$).

The electron–nuclear spin entanglement generated here could also be mapped into an entangled state between nuclear spin pairs[25]. By interchanging (by SWAP) the state of the electron spin with a second, coupled nucleus, for example, nuclear spin entanglement could be attained in a regime where the thermal polarization of the nuclei would be orders of magnitude too small and the direct coupling between them weak. Clusters of up to eight nuclei coupled to a single electron spin have been explored in other materials[26], although the scaling of such an approach seems limited. A scalable network of entangled nuclear spins could be generated by exploiting the ability to ionize the donor and transfer the electron onto a neighbouring donor site[27,28]. These operations, combined with single-shot read-out of the phosphorus donor spin[29] and globally controlled electron–nuclear spin entanglement such as we have demonstrated, form the basis for a cluster-state quantum computer in silicon[25].

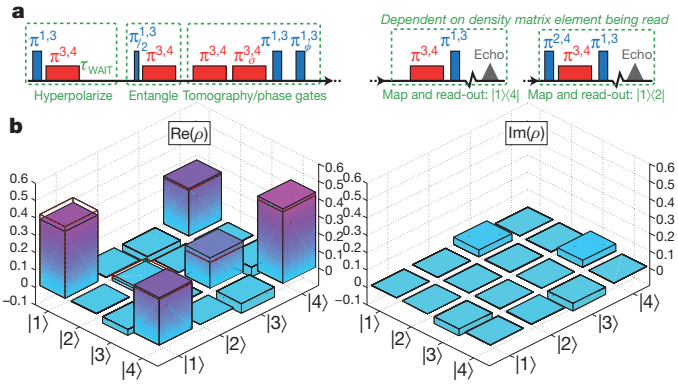


Figure 3 | Measuring an entangled density matrix. a, The full pulse sequence used to prepare, entangle and measure the two-spin state. The final read-out stage was changed according to the density matrix element being measured: examples are shown for the $|1\rangle\langle 2|$ and $|1\rangle\langle 4|$ states. **b**, The obtained density matrix is shown as solid bars, and the dashed outline (zero where not shown) shows that of an ideal state given $\alpha = 0.217$. The fidelity of the ideal state with the measured density matrix is 98%.

METHODS SUMMARY

Si:P consists of an electron spin, $S = 1/2$ ($g = 1.9987$), coupled to the nuclear spin, $I = 1/2$, of ^{31}P through an isotropic hyperfine coupling of $a = 4.19$ mT. The W-band electron spin resonance signal comprises two lines (one for each nuclear spin projection, $M_I = \pm 1/2$). Our experiments were performed on the low-field line of the electron spin resonance doublet, corresponding to $M_I = 1/2$. At 2.9 K and 3.36 T, the electron and nuclear spin relaxation times were measured to be approximately 0.6 s and 100 s, respectively.

The sample consists of a ^{28}Si-enriched single crystal about 0.5 mm in diameter with a residual ^{29}Si concentration of order 70 p.p.m., produced by decomposing isotopically enriched silane in a recirculating reactor to produce poly-silicon rods, followed by floating-zone crystallization. Phosphorus doping of $\sim 10^{14}$ cm^{-3} was achieved by adding dilute PH_3 gas to the ambient argon during the final floating-zone single-crystal growth. Further information on the sample growth has been reported elsewhere[30].

Pulsed electron spin resonance experiments were performed using a W-band (94-GHz) Bruker ELEXSYS 680 spectrometer, modified to allow microwave phase control and equipped with a 6-T superconducting magnet and a low-temperature helium-flow cryostat (Oxford CF935). The cryostat was pumped to achieve a temperature of 2.88 K (internal thermocouple) consistent with the spin temperature measurement (see text). Typical pulse times were 56 ns for a microwave π pulse and 100 μs for a radio-frequency π pulse. To achieve arbitrary phase control, we generated radio-frequency pulses using a Rohde and Schwarz AFQ100B together with an Amplifier Research 500 W amplifier.

Received 30 September; accepted 23 November 2010.
Published online 19 January 2011.

1. Briegel, H.-J., Dür, W., Cirac, J. I. & Zoller, P. Quantum repeaters: the role of imperfect local operations in quantum communication. *Phys. Rev. Lett.* **81,** 5932–5935 (1998).
2. Jozsa, R. & Linden, N. On the role of entanglement in quantum-computational speed-up. *Proc. R. Soc. Lond. A* **459,** 2011–2032 (2003).
3. Curty, M., Lewenstein, M. & Lütkenhaus, N. Entanglement as a precondition for secure quantum key distribution. *Phys. Rev. Lett.* **92,** 217903 (2004).
4. Vandersypen, L. *et al.* Experimental realization of Shor's quantum factoring algorithm using nuclear magnetic resonance. *Nature* **414,** 883–887 (2001).
5. Negrevergne, C. *et al.* Benchmarking quantum control methods on a 12-qubit system. *Phys. Rev. Lett.* **96,** 170501 (2006).
6. Knill, E., Chuang, I. & Laflamme, R. Effective pure states for bulk quantum computation. *Phys. Rev. A* **57,** 3348–3363 (1998).
7. Horodecki, M., Horodecki, P. & Horodecki, R. Separability of mixed states: necessary and sufficient conditions. *Phys. Lett. A* **223,** 1–8 (1996).
8. Peres, A. Separability criterion for density matrices. *Phys. Rev. Lett.* **77,** 1413–1415 (1996).
9. Mehring, M., Mende, J. & Scherer, W. Entanglement between an electron and a nuclear spin 1/2. *Phys. Rev. Lett.* **90,** 153001 (2003).
10. Anwar, M. *et al.* Preparing high purity initial states for nuclear magnetic resonance quantum computing. *Phys. Rev. Lett.* **93,** 040501 (2004).
11. Morton, J. J. L. *et al.* Solid-state quantum memory using the ^{31}P nuclear spin. *Nature* **455,** 1085–1088 (2008).
12. Tyryshkin, A. M. & Lyon, S. A. Data presented at the Silicon Qubit Workshop, 23-24 August (Albuquerque, sponsored by Lawrence Berkeley National Laboratory and Sandia National Laboratory, 2010).
13. Tyryshkin, A. M. *et al.* Coherence of spin qubits in silicon. *J. Phys. Condens. Matter* **18,** S783–S794 (2006).
14. Morton, J. J. L. *et al.* High fidelity single qubit operations using pulsed electron paramagnetic resonance. *Phys. Rev. Lett.* **95,** 200501 (2005).
15. Kane, B. E. A silicon-based nuclear spin quantum computer. *Nature* **393,** 133–137 (1998).
16. Wollan, D. S. Dynamic nuclear polarization with an inhomogeneously broadened ESR line. II. Experiment. *Phys. Rev. B* **13,** 3686–3696 (1976).
17. Hayashi, H., Itahashi, T., Itoh, K. M., Vlasenko, L. S. & Vlasenko, M. P. Dynamic nuclear polarization of ^{29}Si nuclei in isotopically controlled phosphorus doped silicon. *Phys. Rev. B* **80,** 045201 (2009).
18. Schulman, L., Mor, T. & Weinstein, Y. Physical limits of heat-bath algorithmic cooling. *Phys. Rev. Lett.* **94,** 120501 (2005).
19. Wei, T. *et al.* Maximal entanglement versus entropy for mixed quantum states. *Phys. Rev. A* **67,** 022110 (2003).
20. Verstraete, F., Audenaert, K. & De Moor, B. Maximally entangled mixed states of two qubits. *Phys. Rev. A* **64,** 012316 (2001).
21. Aharonov, Y. & Anandan, J. Phase change during a cyclic quantum evolution. *Phys. Rev. Lett.* **58,** 1593–1596 (1987).
22. Suter, D., Mueller, K. T. & Pines, A. Study of the Aharonov-Anandan quantum phase by NMR interferometry. *Phys. Rev. Lett.* **60,** 1218–1220 (1988).

23. Simmons, S., Jones, J. A., Karlen, S. D., Ardavan, A. & Morton, J. J. L. Magnetic field sensors using 13-spin cat states. *Phys. Rev. A* **82,** 022330 (2010).
24. Campbell, E. T. Distributed quantum-information processing with minimal local resources. *Phys. Rev. A* **76,** 040302 (2007).
25. Morton, J. J. L. A silicon-based cluster state quantum computer. Preprint at ⟨http://128.84.158.119/abs/0905.4008v1⟩ (2009).
26. Mehring, M. & Mende, J. Spin-bus concept of spin quantum computing. *Phys. Rev. A* **73,** 052303 (2006).
27. Skinner, A., Davenport, M. & Kane, B. Hydrogenic spin quantum computing in silicon: a digital approach. *Phys. Rev. Lett.* **90,** 087901 (2003).
28. Andresen, S. *et al.* Charge state control and relaxation in an atomically doped silicon device. *Nano Lett.* **7,** 2000–2003 (2007).
29. Morello, A. *et al.* Single-shot readout of an electron spin in silicon. *Nature* **467,** 687–691 (2010).
30. Becker, P., Pohl, H.-J., Riemann, H. & Abrosimov, N. Enrichment of silicon for a better kilogram. *Phys. Status Solidi. A* **207,** 49–66 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Fitzsimons, S. Benjamin, A. Ardavan, A. Briggs and B. Lovett for discussions, and P. Höfer and Bruker BioSpin for support with instrumentation. Three-dimensional images were created using POV-RAY open-source software. We thank EPSRC for supporting work at Oxford through CAESR (EP/D048559/1) and the Oxford–Keio collaboration through the JST-EPSRC SIC programme (EP/H025952/1). Work at Keio has been supported by Grants-in-aid for Scientific Research by MEXT, FIRST by JSPS, Nanoquine and Keio GCOE. S.S. is supported by the Clarendon Fund, J.J.L.M. is supported by St John's College, Oxford, and the Royal Society.

Author Contributions S.S., R.M.B. and J.J.L.M. designed and performed the experiments and wrote the paper. H.R., N.V.A., P.B. and H.-J.P. grew the ^{28}Si crystal. K.M.I and M.L.W.T. analysed and prepared the sample and discussed the experiments, results and manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.J.L.M. (john.morton@materials.ox.ac.uk).

CORRIGENDUM

doi:10.1038/nature09712

Branched tricarboxylic acid metabolism in *Plasmodium falciparum*

Kellen L. Olszewski, Michael W. Mather, Joanne M. Morrisey, Benjamin A. Garcia, Akhil B. Vaidya, Joshua D. Rabinowitz & Manuel Llinás

Nature **466,** 774–778 (2010)

The samples used for histone proteomics described in this Letter were inadvertently switched, such that the U-^{13}C-glucose and U-^{13}C-^{15}N-glutamine data were inverted. The plots in Fig. 2b and the spectra in Supplementary Fig. 3 have been modified to reflect this. The corrected results demonstrate that ^{13}C-labelling of histone acetyl groups occurs only in cells grown on ^{13}C-glucose and not on ^{13}C-glutamine. Therefore, glucose is the primary source of the acetyl units used for both amino sugar biosynthesis and nuclear protein acetylation. Although U-^{13}C-^{15}N-glutamine does give rise to labelled acetyl-CoA, its localization and function remain unclear. The model presented in Fig. 4 has been modified to reflect these facts, which do not alter the paper's main conclusions about TCA cycle architecture. The corrected Figs 2b and 4 are shown below. The authors apologize for this error.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

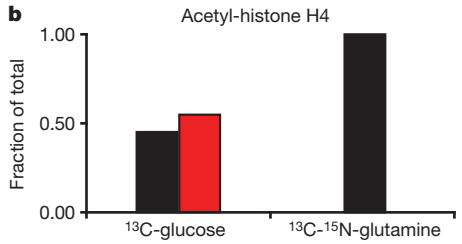


Figure 2

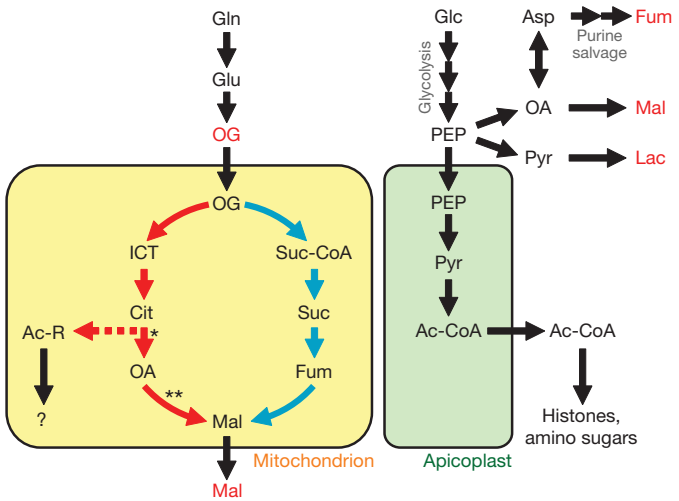


Figure 4

LETTER

doi:10.1038/nature09747

Modelling the long QT syndrome with induced pluripotent stem cells

Ilanit Itzhaki[1]*, Leonid Maizels[1]*, Irit Huber[1]*, Limor Zwi-Dantsis[1], Oren Caspi[1], Aaron Winterstern[1], Oren Feldman[1], Amira Gepstein[1], Gil Arbel[1], Haim Hammerman[2], Monther Boulos[2] & Lior Gepstein[1,2]

The ability to generate patient-specific human induced pluripotent stem cells (iPSCs)[1–3] offers a new paradigm for modelling human disease and for individualizing drug testing[4]. Congenital long QT syndrome (LQTS) is a familial arrhythmogenic syndrome characterized by abnormal ion channel function and sudden cardiac death[5–7]. Here we report the development of a patient/disease-specific human iPSC line from a patient with type-2 LQTS (which is due to the A614V missense mutation in the *KCNH2* gene). The generated iPSCs were coaxed to differentiate into the cardiac lineage. Detailed whole-cell patch-clamp and extracellular multielectrode recordings revealed significant prolongation of the action-potential duration in LQTS human iPSC-derived cardiomyocytes (the characteristic LQTS phenotype) when compared to healthy control cells. Voltage-clamp studies confirmed that this action-potential-duration prolongation stems from a significant reduction of the cardiac potassium current I_{Kr}. Importantly, LQTS-derived cells also showed marked arrhythmogenicity, characterized by early-after depolarizations and triggered arrhythmias. We then used the LQTS human iPSC-derived cardiac-tissue model to evaluate the potency of existing and novel pharmacological agents that may either aggravate (potassium-channel blockers) or ameliorate (calcium-channel blockers, K_{ATP}-channel openers and late sodium-channel blockers) the disease phenotype. Our study illustrates the ability of human iPSC technology to model the abnormal functional phenotype of an inherited cardiac disorder and to identify potential new therapeutic agents. As such, it represents a promising paradigm to study disease mechanisms, optimize patient care (personalized medicine), and aid in the development of new therapies.

Congenital LQTS[5–7], now classified into 12 subtypes, is a familial arrhythmogenic syndrome characterized by delayed repolarization, a prolonged QT interval in the electrocardiogram and a life-threatening polymorphic ventricular tachycardia known as torsade de pointes (TdP). Although heterologous expression systems[6] and animal models have provided important insights into LQTS pathogenesis, the lack of *in vitro* sources for human cardiomyocytes and the inability to model patient-specific disease variations has significantly hampered the study of this disease. Here we propose that cardiomyocytes derived from LQTS patient-specific human iPSC lines may recapitulate the disease phenotype *in vitro*, providing important mechanistic insights and offering a unique platform to evaluate patient-specific disease aggravators and therapies (Supplementary Fig. 1).

Dermal fibroblasts were obtained from a 28-year-old woman with a diagnosis of familial type-2 LQTS (Fig. 1a) due to a missense mutation in exon 9 of the *KCNH2* gene (A614V mutation caused by a C→T nucleotide substitution)[8]. This mutation was previously shown in heterologous expression systems to affect the pore-forming region of the *KCNH2* (also known as *HERG*)-encoded potassium channel (Supplementary Fig. 1); leading to a significant reduction of the rapid component of the delayed-rectifier potassium current (I_{Kr}) responsible for type-2 LQTS[9].

The cultured fibroblasts were reprogrammed to generate LQTS patient-specific human iPSC clones (Fig. 1b) after transduction with retroviral vectors encoding for SOX2, KLF4 and OCT4. Several clones were generated, three of which were continuously propagated and used for cardiomyocyte differentiation and characterization. Similar control human iPSCs were created from fibroblasts from a healthy individual. Importantly, the A614V heterozygous mutation was identified in LQTS human iPSCs (Fig. 1a) but not in control iPSCs.

All iPSC clones generated showed characteristic human embryonic stem (ES) cell morphology (Fig. 1b and Supplementary Fig. 2), expressed the pluripotency markers NANOG, SSEA4, OCT4 and TRA-1-60 (Fig. 1b and Supplementary Fig. 2), had alkaline phosphotase activity (Fig. 1b and Supplementary Fig. 3A), and maintained a normal karyotype (Fig. 1c and Supplementary Fig. 3B). All clones showed silencing of the three retroviral transgenes (Supplementary Fig. 4), reactivation of endogenous pluripotency genes (*OCT4*, *SOX2*, *NANOG*, *FOXD3* and *ZFP42* (also known as *REX1*); Supplementary Fig. 5), and demethylation of the *NANOG* promoter regions (Fig. 1d); all indicating successful reprogramming. Pluripotency of the human iPSC clones was confirmed by the presence of cell derivatives of all three germ layers in *in-vitro*-differentiating embryoid bodies (Fig. 1e and Supplementary Fig. 6) and by teratoma formation after injection of undifferentiated human iPSCs into immunocompromised NOD/SCID mice (Fig. 1f and Supplementary Fig. 7).

Next we used the embryoid-body-differentiation system to coax human iPSC differentiation into the cardiac lineage. Similarly to human ES cells[10,11] and healthy human iPSCs[12,13], this resulted in the appearance of spontaneously beating areas within the differentiating embryoid bodies (Supplementary Movies 1 and 2). Both the control and LQTS human iPSC lines showed comparable cardiomyocyte differentiation capacities (Supplementary Fig. 8). Immunocytostaining studies showed positive staining of the control (Supplementary Fig. 9A) and LQTS human iPSC-derived cardiomyocytes (iPSC-CMs) for the sarcomeric proteins troponin I (TnI) and α-actinin and for the gap-junction protein connexion 43 (Fig. 2a). Similarly, gene expression analysis confirmed the expression of cardiac-specific transcription factors (*NKX2-5*) and structural genes (*MLC2V* (also known as *MYL2*), *MYH6* and *MYH7*) by both cell types, including of the relevant *KCNH2* gene (Fig. 2b). Importantly, both control and LQTS iPSC-CMs showed cardiac-specific functional properties. This was manifested by the development of action potentials at the cellular level (Fig. 3a), a functional syncytium at the multicellular level (Fig. 2c and Supplementary Fig. 9B) and appropriate chronotropic responses to the β-adrenergic agonist isoproterenol (Fig. 2d and Supplementary Fig. 9C).

Similar to previous reports in human ES cells[10] and human iPSCs[12], three types of action-potential morphologies were recorded from control and LQTS iPSC-CMs (Fig. 3a and Supplementary Table 1). The most dominant form was the 'ventricular-like' morphology (characterized by a significant plateau phase), with the 'atrial-like' (showing

[1]Sohnis Family Research Laboratory for Cardiac Electrophysiology and Regenerative Medicine, the Bruce Rappaport Faculty of Medicine, Technion — Israel Institute of Technology, POB 9649, Haifa 31096, Israel. [2]Department of Cardiology, Rambam Medical Center, the Bruce Rappaport Faculty of Medicine, Technion — Israel Institute of Technology, POB 9649, Haifa 31096, Israel.
*These authors contributed equally to this work.

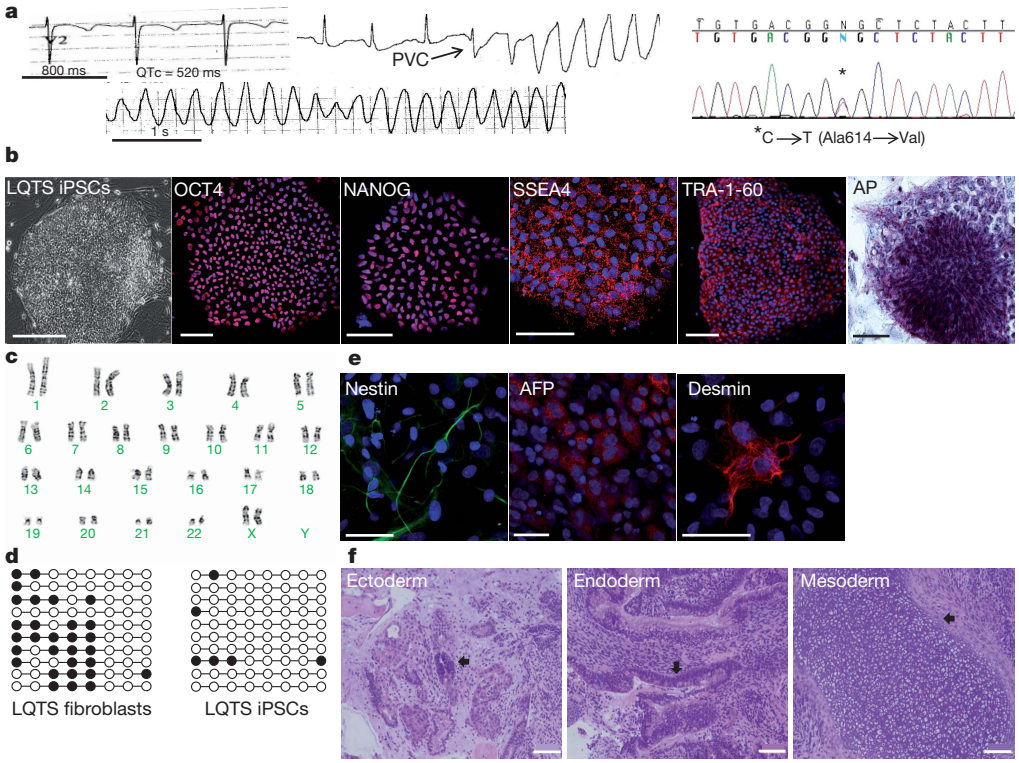


Figure 1 | Establishment and characterization of type-2 LQTS human iPSCs (clone 1). **a**, Left panel shows ECGs from the type-2 LQTS patient during sinus rhythm (QT interval corrected for heart rate (QTc) = 520 ms) and during initiation and sustainment of TdP. PVC, premature ventricular contraction. Right panel shows sequencing of the *KCNH2* gene identifying the A614V heterozygous missense mutation in LQTS human iPSCs. **b**, Reprogramming of the patient's fibroblasts into LQTS human iPSCs. Shown from left to right are a typical human iPSC colony, positive immunostaining for pluripotency markers (OCT4, NANOG, SSEA4 and TRA-1-60), and positive staining for alkaline phosphatase (AP) activity. Scale bars, 100 μm. **c**, Karyotype analysis of LQTS human iPSCs. **d**, Bisulphite sequencing analysis of the NANOG promoter in the patient's fibroblasts and LQTS human iPSCs. **e**, Immunostaining of *in vitro* differentiating embryoid bodies for nestin (ectoderm), α-fetoproptein (AFP, endoderm) and desmin (mesoderm). Scale bars, 50 μm. **f**, Teratoma formation following injection of undifferentiated LQTS human iPSCs in NOD/SCID mice. Note the formation of pigmented epithelium (ectoderm), gastrointestinal epithelium (endoderm) and hyaline cartilage (mesoderm), as identified by the arrows. Scale bars, 200 μm.

triangular-shaped action potentials) and 'nodal-like' (showing relatively depolarized resting membrane potential) types being less prevalent. More importantly, these intracellular recordings revealed marked action-potential duration (APD) prolongation in LQTS iPSC-CMs (the characteristic LQTS electrophysiological signature[5–7]) when compared to control cells (Fig. 3a, b and Supplementary Table 1). APD prolongation and the associated reduced repolarization velocity were noted in both ventricular-like and atrial-like LQTS human iPSC-CMs but not in the nodal-like cells.

Because APD is rate-dependent, next we electrically stimulated the human iPSC-CMs and noted that the differences in APD values between LQTS and control cells persisted also at fixed rates (0.5 and 1 Hz; Fig. 3b and Supplementary Fig. 10). In addition, comparable prolonged APD values were noted in ventricular myocytes derived from three different independent LQTS iPSC clones, which were all significantly longer than those obtained from two different control human iPSC lines (Supplementary Fig. 11A). Lastly, the LQTS phenotype could be recapitulated in control iPSC-CMs by pharmacologically blocking the I_{Kr} current with the specific blocker E-4031 (Supplementary Fig. 12).

Next we performed single-cell voltage-clamp studies and identified the presence of an E4031-sensitive current (I_{Kr}) in control human iPSC-CMs (Fig. 3c), which resembled previous I_{Kr} recordings from explanted human ventricular myocytes[14]. Similar recordings in the LQTS iPSC-CMs revealed significant reductions of this current (Fig. 3c). Thus, peak amplitudes of the I_{Kr} activation currents in LQTS cardiomyocytes ($n = 6$), measured at physiologically relevant depolarization steps (0, 20 mV), were found to be significantly smaller (by 72 and 60%, respectively) than those recorded in control cells ($P < 0.05$; Fig. 3d, left). Similarly, peak amplitudes of the I_{Kr} tail currents measured at 20 and 60 mV were also significantly smaller (by 64 and 68%, respectively, $P < 0.05$; Fig. 3d, right) than in control cells. These results are consistent with previous recordings in heterologous expression systems, where co-injection of wild-type along with mutant A614V cRNA suppressed the HERG current in a dominant-negative manner[9].

To evaluate the electrophysiological properties also at the multicellular level, we studied the human iPSC-derived cardiac tissue with a microelectrode array mapping technique[15]. The recorded extracellular electrograms were analysed to measure the field-potential duration (FPD; Fig. 2e). This measurement, which was normalized to account for variations in beating frequency[16] (corrected FPD, cFPD), is analogous to the QT interval in the electrocardiogram (ECG) and was previously shown to correlate with APD and to predict drug effects on repolarization[15,17]. Similar to patch-clamp studies, cFPD (or rate-matched FPD) measurements were significantly longer in cardiac tissues derived from all three LQTS human iPSC clones when compared to healthy control specimens (Fig. 2e, f, Supplementary Fig. 11B and Supplementary Movie 1).

Importantly, LQTS iPSC-CMs showed marked arrhythmogenicity. This was manifested by the development of early-after depolarizations (EADs) in both the atrial-like and ventricular-like LQTS iPSC-CMs (Fig. 3e). EADs are spontaneous membrane depolarizations regarded as the harbinger of ventricular arrhythmias in LQTS[7], and are thought to result from late membrane-inward currents, such as the L-type Ca^{2+} 'window' current[18]. When reaching threshold, EADs may give rise to triggered action potentials, manifested in LQTS iPSC-CMs as isolated premature beats or even multiple sequential triggered action potentials (Fig. 3f, g). The presence of EADs was observed in 38 of the 58 LQTS iPSC-CMs studied (66%). These EADs gave rise to a single premature

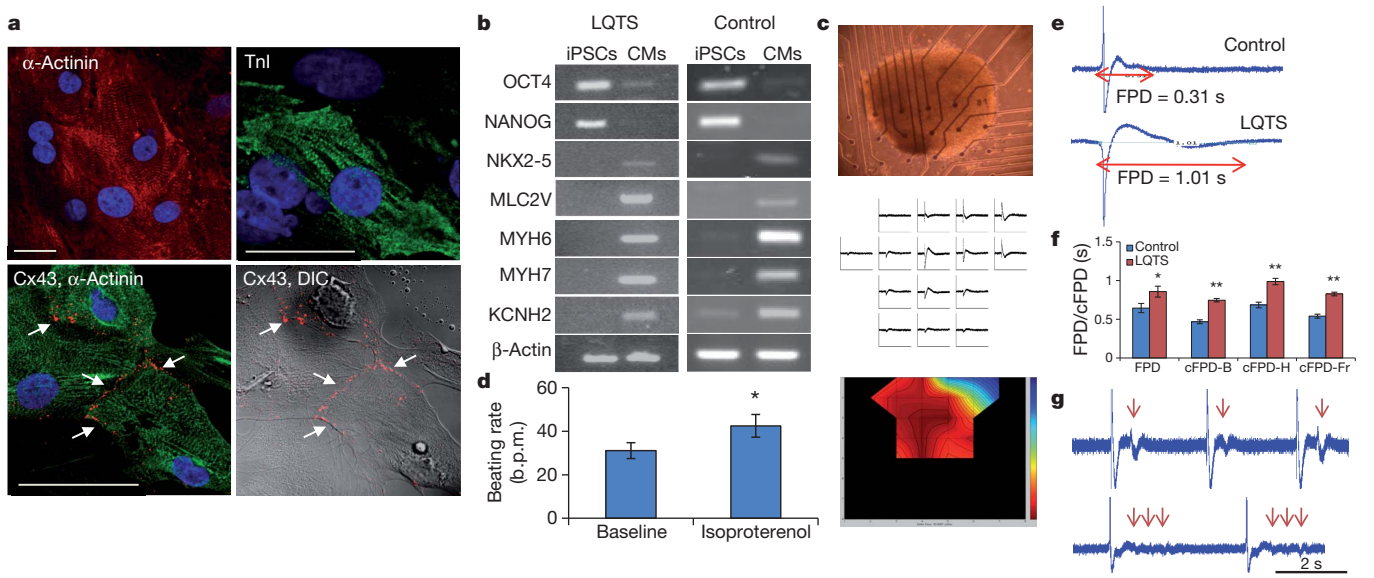


Figure 2 | Phenotypic characterization of LQTS human iPSC-derived cardiac-tissue. **a**, Immunostaining of LQTS iPSC-CMs for α-actinin and TnI (top). Gap junctions (arrows) were identified by the punctuated connexion 43 (Cx43) immunosignal (red) at the intercellular junctions (differential interference contrast microscopy (DIC) image, bottom right) between the cardiomyocytes (identified by α-actinin staining in green, bottom left). Scale bars, 30 μm. **b**, Semi-quantitative polymerase chain reaction with reverse transcription (RT–PCR) demonstrating expression of cardiac-specific transcription factors (*NKX2-5*), sarcomeric proteins (*MYH6, MYH7, MLC2V*) and ion channels (*KCNH2*) by LQTS and control human iPSC-CMs in comparison to undifferentiated human iPSCs expressing *OCT4* and *NANOG*. **c**, Microelectrode array extracellular recordings (middle) from LQTS human iPSC cardiac tissue (top). The resulting colour-coded activation map (right) shows electrical propagation from the pacemaker area (red) to latest activation sites (blue). **d**, Positive chrontropic responses of LQTS hiPSCs-CMs ($n = 16$) to isoproterenol (10μM). $^*P < 0.01$ when compared to baseline. b.p.m., beats per minute. **e**, Extracellular recordings from control (top) and LQTS (bottom) human iPSC-derived cardiac tissues and their respective FPD measurements. **f**, Summary of rate-matched FPD ($n = 12$) and cFPD values (using Bazett's (B), Hodges' (H) or Fridericia's[16] (Fr) corrections) in control ($n = 23$) and LQTS human iPSC ($n = 45$) cardiac tissues. Error bars show s.e.m. $^*P < 0.05$ and $^{**}P < 0.01$. **g**, Arrhythmogenicity in LQTS cardiac tissue identified as single (top) or multiple (bottom, arrows) premature beats.

beat in 36% of the cells and in 28% also to sustained triggered activity. In contrast, recordings from control iPSC-CMs ($n = 31$) failed to show any EADs or any other arrhythmogenic activity.

The arrhythmogenic activity of the LQTS iPSC-CMs was also identified at the multicellular level and ranged from isolated premature beats (Fig. 2g, top, and Supplementary Movie 2) to more sustained activity (Fig. 2g, bottom). Consequently, ectopic activity in the microelectrode array recordings was noted in 38% (26 of 84) of the LQTS human iPSC-derived cardiac tissues studied but only in 6% (3 of 47) of control samples. This ectopic activity was consistent with a triggered activity mechanism (although re-entry and automaticity could not be ruled out completely).

Next we used the LQTS human iPSC model to evaluate drugs that may either ameliorate or aggravate the disease phenotype[19]. Because malignant arrhythmias in LQTS patients are often precipitated by drugs that block I_{Kr}, we initially evaluated the specific I_{Kr} blocker E-4031 (500 nM), and noted significant APD prolongation (by $47 \pm 2\%$, $P < 0.01$, $n = 7$) in the LQTS iPSC-CMs (Supplementary Fig. 12B). This APD prolongation was coupled with increased arrhythmogenesis in 6 of the 7 cells studied, manifested either by the development of new EADs or by an increase in their number and complexity (Supplementary Fig. 12B). Similarly, I_{Kr} blockade by E-4031 or by cisapride (a gastric prokinetic agent previously removed from the market because of its indirect I_{Kr} blocking activity, which led to increased pro-arrhythmic mortality[20]) significantly prolonged cFPD and increased arrhythmogenicity also at the multicellular level (Supplementary Fig. 12C, D). These results, together with the I_{Kr} current analysis, provide a measure of the magnitude of the dominant-negative effect induced by the A614V mutation. They also offer clues to the potential susceptibility of such LQTS patients to TdP in response to agents that may further inhibit the remaining I_{Kr} current. More broadly it highlights the potential role of the human iPSC model for safety pharmacology[20], both for drug development and for patient-specific safety screening.

Next we used the LQTS human iPSC-CM model to study agents that may possess novel therapeutic benefits for LQTS, through either APD shortening or by direct EAD suppression[19]. The effect of these agents was also evaluated in healthy control cells (Supplementary Fig. 13). Because Ca^{2+} influx through L-type Ca^{2+} channels contributes importantly to APD and also has a role in EAD formation[18], we proposed that inhibition of this current may be anti-arrhythmic[19]. Application of a Ca^{2+}-channel blocker (nifedipine, 1 μM) led to significant APD abbreviation in LQTS cardiomyocytes (Fig. 4a, b, top), with mean APD_{90} (the time interval required to reach 90% of repolarization) shortened by $57 \pm 7\%$ ($P < 0.01$). More importantly, nifedipine application resulted in complete elimination of all EADs and triggered beats (Fig. 4c; $n = 11$). Similarly, nifedipine (1 μM) significantly shortened cFPD (Fig. 4a; $P < 0.01$, $n = 5$) and abolished all arrhythmic activity in the LQTS cardiac-tissue model (Fig. 4d). Long-term application of this agent, however, was associated with cessation of beating in some embryoid bodies.

An alternative strategy to favourably alter the balance between inward and outward currents during repolarization in the LQTS may be to augment outward potassium currents. To this end we evaluated the effects of pinacidil (1 μM), a K_{ATP}-channel opener, and noted significant shortening of both APD_{90} (Fig. 4a, b; $P < 0.05$, $n = 4$) and cFPD (Fig. 4a; $P < 0.05$, $n = 9$) in LQTS iPSC-CMs. Importantly, pinacidil application completely abolished all EADs and triggered arrhythmias in all LQTS iPSC-CMs studied (Fig. 4e, f; $n = 7$). A potential shortcoming of the aforementioned therapies may be excessive APD shortening leading to the potentially pro-arrhythmic 'short QT syndrome'[19]. In this respect, the human iPSC model may prove useful, as it may allow the identification of compounds and optimization of their dosage to result in APD normalization without pathological shortening.

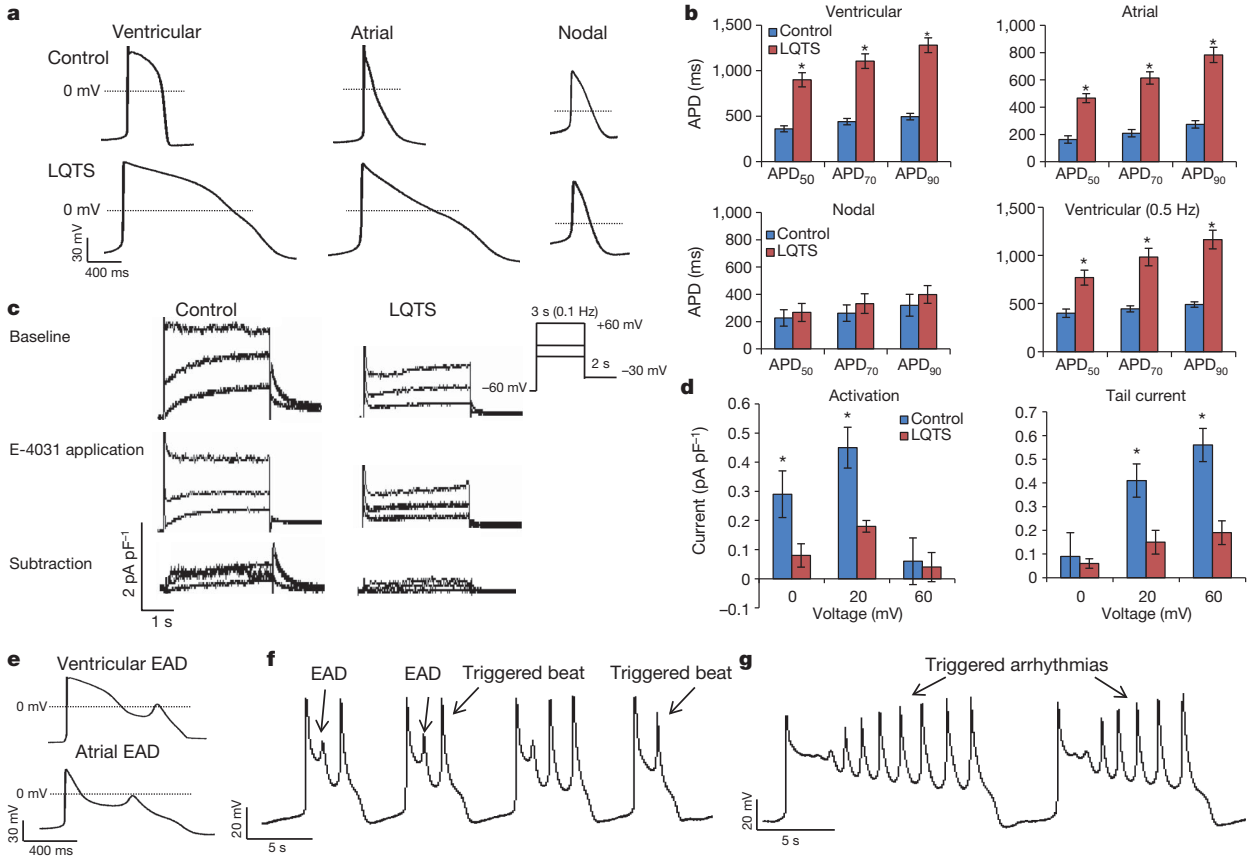


Figure 3 | Whole-cell patch-clamp recording from human iPSC-CMs. **a**, Action-potential recordings from control and LQTS human iPSC-CMs showing ventricular-like, atrial-like and nodal-like morphologies. Notice the marked APD prolongation in both ventricular-like and atrial-like LQTS iPSC-CMs. **b**, APD_{90}, APD_{70}, and APD_{50} values in LQTS (red) and healthy (blue) human iPSC-CMs. $*P < 0.01$ when compared to control cells. Error bars show s.e.m. **c**, Voltage-clamp recordings of the I_{Kr} current, measured from control (left) and LQTS (right) human iPSC-derived ventricular cardiomyocytes. Top, baseline recordings. Middle, recording following administration of E-4031 (1 μM). Bottom, E-4031-sensitive current (I_{Kr}) defined by digital subtraction of the two currents. **d**, Summary of the I_{Kr} activation and tail-current amplitudes measured following test depolarization pulses of 0, 20 and 60 mV. Note the significant I_{Kr} attenuation in LQTS cells (red, $n = 6$) when compared to controls (blue, $n = 6$). Error bars show s.e.m. **e**, Development of EADs in LQTS ventricular-like and atrial-like cells. **f**, **g**, Development of triggered activity in LQTS cardiomyocytes, manifested as EADs or even single and multiple (**g**) triggered beats.

Lastly, we also evaluated the potential anti-arrhythmic efficacy of the late Na^+-channel blocker, ranolazine[21]. Although, this agent may theoretically be more applicable for type-3 LQTS (resulting from gain-of-function Na^+-channel mutations), a recent study showed its beneficial effects also in a canine model of acquired type-2 LQTS[22]. Interestingly, ranolazine application (15–50 μM) did not significantly alter APD_{90} or cFPD in LQTS cardiomyocytes (Fig. 4a and Supplementary Fig. 14; $n = 8$), probably because of its nonspecific blocking effect on different ion channels. Nevertheless, ranolazine induced a pronounced anti-arrhythmic effect at both the cellular and multicellular levels (Fig. 4g, h and Supplementary Fig. 14).

Taken together, our data demonstrate the ability of human iPSC technology to model the abnormal functional phenotype of an inherited cardiac disorder. Using this approach, the disease phenotype presented by type-2 LQTS patients at the bedside (long QT interval and the development of TdP) could be recognized as a measurable anomalous electrophysiological signature in bench studies using patient-specific human iPSC-CMs. Our study adds to a recent report demonstrating the ability to model type-1 LQTS using human iPSCs[23]. Importantly, the current study provides mechanistic insights into the pathogenesis of arrhythmias in this syndrome (by demonstrating the development of spontaneous EADs leading to triggered activity at both the cellular and multicellular levels). Moreover, it also highlights the potential of human iPSCs in the emerging field of personalized medicine by demonstrating the ability to screen the effects of potential disease aggravators and novel customized treatment options (which in LQTS include Ca^{2+}-channel blockers, K_{ATP}-channel openers, and late Na^+-channel blockers).

The results of the current study, focusing on congenital type-2 LQTS, may have broader implications, as the specific current affected (I_{Kr}) is thought to have an important role in the more commonly acquired LQTS forms (resulting from heart failure, cardiac hypertrophy, or drug therapy). Moreover, the concepts described here may also be extended to model several other human genetic disorders, enabling translational research into disease mechanisms and therapies.

METHODS SUMMARY

Human iPSC generation and cardiomyocyte differentiation. Dermal fibroblasts were reprogrammed to generate human iPSC clones by transduction with retroviral vectors encoding SOX2, KLF4 and OCT4, followed by valproic acid treatment[24]. The undifferentiated human iPSC colonies were propagated on a MEF feeder layer. To induce differentiation, human iPSCs were dispersed into small cell clumps and cultivated in suspension for 10 days as embryoid bodies and then plated[11,13].

Immunostainings. Cells were fixed with 4% paraformaldehyde, permeabilized with 1% Triton, blocked with 5% horse serum, and incubated with primary antibodies (as listed in Methods). The preparations were incubated with secondary antibodies and counterstained with 4′,6-diamidino-2-phenylindole (DAPI) for visualization.

Whole-cell patch-clamp recordings. Dispersed human iPSC-CMs, obtained by enzymatic dissociation (collagenase, 1 mg ml^{-1}) of contracting embryoid bodies, were attached to fibronectin-coated glass coverslips. Action potentials were recorded at 32 °C in the current-clamp mode. A custom-written Matlab program was used to analyse action-potential parameters. Voltage-clamp studies were performed to characterize the I_{Kr} current, defined as the E-4031 (1 μM) sensitive current. These

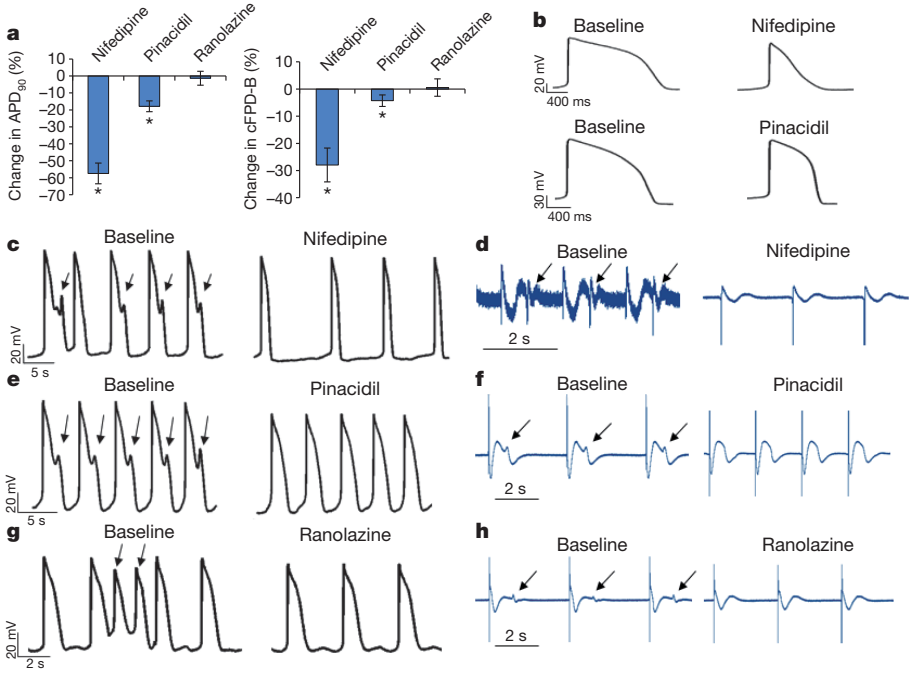


Figure 4 | Drug screening using LQTS human iPSC-CMs. **a**, Changes (percentage) in APD_{90} and cFPD with Bazett's correction values following treatment with Ca^{2+}-channel blocker nifedipine, K_{ATP}-channel opener pinacidil and late Na^{+}-channel blocker ranolazine. $*P < 0.05$ when mean absolute values are compared to baseline. Error bars show s.e.m. **b**, APD shortening induced by nifedipine and pinacidil in LQTS iPSC-CMs. **c**–**h**, Anti-arrhythmic activity at the cellular (**c**, **e**, **g**) and multicellular (**d**, **f**, **h**) levels induced by nifedipine (**c**, **d**), pinacidil (**e**, **f**) and ranolazine (**g**, **h**). Arrows show the ectopic activity at baseline.

currents were elicited by 3-s depolarizing steps from a holding potential of −60 mV to potentials ranging from −40 to +60 mV in 10 mV increments. This was followed by a 2-s repolarization phase to −30 mV to elicit the tail current.

Microelectrode array recordings. Extracellular electrograms, recorded by a high-resolution microelectrode array recording system (multichannel systems)[15], were used to determine FPD values in LQTS and control human iPSC-derived cardiac tissues. FPD was defined as the time interval between the initial field-potential deflection and its return to baseline. FPD measurements were either compared in rate-matched specimens or normalized (cFPD) for rate using Bazett's, Hodges' or Fridericia's[16] correction formulas.

Statistical analysis. Results are reported as mean ± s.e.m. Comparisons between LQTS human iPSC-CMs and healthy human iPSC-CMs groups were performed using the unpaired Student's *t*-test. Drug effects were analysed by paired *t*-test. One-way ANOVA followed by Tukey's post-hoc tests were used when comparing multiple groups. $P < 0.05$ was considered statistically significant.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 June; accepted 14 December 2010.
Published online 16 January 2011.

1. Park, I. H. *et al.* Disease-specific induced pluripotent stem cells. *Cell* **134,** 877–886 (2008).
2. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131,** 861–872 (2007).
3. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318,** 1917–1920 (2007).
4. Kiskinis, E. & Eggan, K. Progress toward the clinical application of patient-specific pluripotent stem cells. *J. Clin. Invest.* **120,** 51–59 (2010).
5. Goldenberg, I. & Moss, A. J. Long QT syndrome. *J. Am. Coll. Cardiol.* **51,** 2291–2300 (2008).
6. Sanguinetti, M. C. & Tristani-Firouzi, M. hERG potassium channels and cardiac arrhythmia. *Nature* **440,** 463–469 (2006).
7. Marbán, E. Cardiac channelopathies. *Nature* **415,** 213–218 (2002).
8. Tenenbaum, M. *et al.* Identification of the gene causing long QT syndrome in an Israeli family. *Isr. Med. Assoc. J.* **10,** 809–811 (2008).
9. Nakajima, T. *et al.* Novel mechanism of HERG current suppression in LQT2: shift in voltage dependence of HERG inactivation. *Circ. Res.* **83,** 415–422 (1998).
10. He, J. Q., Ma, Y., Lee, Y., Thomson, J. A. & Kamp, T. J. Human embryonic stem cells develop into multiple types of cardiac myocytes: action potential characterization. *Circ. Res.* **93,** 32–39 (2003).
11. Kehat, I. *et al.* Human embryonic stem cells can differentiate into myocytes with structural and functional properties of cardiomyocytes. *J. Clin. Invest.* **108,** 407–414 (2001).
12. Zhang, J. *et al.* Functional cardiomyocytes derived from human induced pluripotent stem cells. *Circ. Res.* **104,** e30–e41 (2009).
13. Zwi, L. *et al.* Cardiomyocyte differentiation of human induced pluripotent stem cells. *Circulation* **120,** 1513–1523 (2009).
14. Li, G. R., Feng, J., Yue, L., Carrier, M. & Nattel, S. Evidence for two components of delayed rectifier K^{+} current in human ventricular myocytes. *Circ. Res.* **78,** 689–696 (1996).
15. Caspi, O. *et al. In vitro* electrophysiological drug testing using human embryonic stem cell derived cardiomyocytes. *Stem Cells Dev.* **18,** 161–172 (2009).
16. Luo, S., Michler, K., Johnston, P. & Macfarlane, P. W. A comparison of commonly used QT correction formulae: the effect of heart rate on the QTc of normal ECGs. *J. Electrocardiol.* **37** (suppl. 1), 81–90 (2004).
17. Halbach, M., Egert, U., Hescheler, J. & Banach, K. Estimation of action potential changes from field potential recordings in multicellular mouse cardiac myocyte cultures. *Cell. Physiol. Biochem.* **13,** 271–284 (2003).
18. January, C. T. & Riddle, J. M. Early afterdepolarizations: mechanism of induction and block. A role for L-type Ca^{2+} current. *Circ. Res.* **64,** 977–990 (1989).
19. Patel, C. & Antzelevitch, C. Pharmacological approach to the treatment of long and short QT syndromes. *Pharmacol. Ther.* **118,** 138–151 (2008).
20. Fermini, B. & Fossa, A. A. The impact of drug-induced QT interval prolongation on drug discovery and development. *Nature Rev. Drug Discov.* **2,** 439–447 (2003).
21. Antzelevitch, C. *et al.* Electrophysiological effects of ranolazine, a novel antianginal agent with antiarrhythmic properties. *Circulation* **110,** 904–910 (2004).
22. Antoons, G. *et al.* Late Na^{+} current inhibition by ranolazine reduces torsades de pointes in the chronic atrioventricular block dog model. *J. Am. Coll. Cardiol.* **55,** 801–809 (2010).
23. Moretti, A. *et al.* Patient-specific induced pluripotent stem-cell models for long-QT syndrome. *N. Engl. J. Med.* **363,** 1397–1409 (2010).
24. Huangfu, D. *et al.* Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. *Nature Biotechnol.* **26,** 795–797 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This study was supported in part by the Israel Science Foundation and Legacy Heritage Foundation (no. 1225/09), by the Yad Hanadiv Foundation Bruno Award, by the Lorry Lokey research fund, and by the Nancy and Stephen Grand Philanthropic Fund. We thank E. Suss-Toby and O. Shenker (from the multidisciplinary laboratory unit) and M. Tzukerman for their valuable help, A. Zamir for writing the MEA analysis software and I. Laevsky and T. Falik-Zaccai for the karyotype analysis.

Author Contributions I.I., L.M., I.H. and L.G. designed the experiments; I.I., L.M., I.H., L.Z.-D., O.C., A.W., O.F., A.G. and G.A performed the experiments; I.I. and L.M. analysed and interpreted the electrophysiological data; M.B. and H.H. performed the clinical assessment; L.G wrote the manuscript; all authors read and approved the manuscript; and L.G. supervised this research work.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to L.G. (mdlior@techunix.technion.ac.il).

METHODS

Establishment and cardiomyocyte differentiation of patient-derived iPSCs. Patient-derived human iPSC clones were established from the patient's skin fibroblasts by retroviral delivery of three reprogramming factors (SOX2, KLF4 and OCT4), followed by application of the histone deacetylase inhibitor valproic acid (VPA)[24]. Briefly, Moloney-based retroviral vectors (pMXs) containing human complementary DNAs (cDNAs) of *OCT4*, *SOX2* and *KLF4* (Addgene plasmids 17964, 17218 and 17219, respectively) were used for retrovirus particle production. These plasmids were cotransfected with the helper plasmid encoding VSVG into HEK-293GP cells for virus production. Virus-containing media were collected at 48 and 72 h after transfection and used for two 24-h rounds of infection of the fibroblasts. Cells were then re-plated at a density of $1–2 \times 10^5$ cells per well on a MEF feeder layer, cultured in ES medium and treated with 0.9 mM VPA for 14 d.

Several human iPSC clones, which morphologically resembled human ES cells and were positively stained with vital TRA-1-81 or TRA-1-60 staining[25], were selected and expanded for further characterization. The undifferentiated human iPSC colonies were cultured on mitotically inactivated mouse embryonic fibroblast (MEF) cells as previously described[11,13,26]. Y-27632 (5 μM), a ROCK inhibitor, was added in early passages to increase seeding efficiency. The culture medium consisted of 80% knockout high-glucose glutamine-free DMEM with sodium pyruvate supplemented with 20% serum replacement, 1 mM L-glutamine, 0.1 mM mercaptoethanol, 4 ng ml^{-1} human recombinant basic fibroblast growth factor (bFGF), and 1%-nonessential amino acid stock (Invitrogen).

Cardiomyocyte differentiation was induced using the embryoid body differentiating system as previously described in detail[13,26]. Undifferentiated human iPSCs were dispersed into small cell clumps using 300 U ml^{-1} of collagenase IV (Worthington) for 45 min. These cells were then cultivated in suspension for 10 days in the differentiation medium (knockout DMEM supplemented with 20% FBS (Hyclone), nonessential amino acids, β-mercaptoethanol and glutamine) where they aggregated to form embryoid bodies. The embryoid bodies were then plated on 0.1% gelatin-coated culture dishes, cultured with the same differentiation medium, and examined daily for the appearance of spontaneous contractions. At 30 days after plating, contracting areas within the embryoid bodies were carefully microdissected and used for the phenotypic characterization studies.

We studied cardiomyocytes derived from three different LQTS human iPSC clones (generated independently from the patient's fibroblasts). They were compared with cardiomyocytes derived from a healthy control human iPSC line generated using the same methods. We also used a second control human iPSC line that was well-characterized previously[13,27].

Genomic sequencing. Genomic DNA was isolated from the patient-derived human iPSC colonies using the high-pure PCR template preparation kit (Roche). The relevant DNA fragment of the *KCNH2* gene was amplified by PCR reaction using 100 ng genomic DNA (primer sequences are detailed in Supplementary Table 2). PCR products were then sequenced.

Immunofluresence and alkaline phosphatase staining. Colonies of undifferentiated human iPSCs and single cardiomyocytes obtained by dissociating beating embryoid bodies were fixed with 4% paraformaldehyde, permeabilized with 1% Triton (Sigma) and blocked with 5% horse serum. Specimens were incubated overnight at 4 °C with primary antibodies targeting OCT4, Tra-1-60, CX43 (from Santa Cruz), NANOG (Peprotech), SSEA4 (R&D), nestin (Millipore), α-fetoprotein (Cell Marque), desmin (Thermo), troponin I (TnI; Chemicon) and sarcomeric α-actinin (Sigma). The preparations were incubated with secondary antibodies at a dilution of 1:200 for 1 h. Nuclei were counterstained with DAPI (1:500, Sigma). Preparations were examined using a laser-scanning confocal microscope (Zeiss LSM-510-PASCAL). Alkaline phosphatase activity was detected in live cultures using the alkaline phosphatase detection kit (Sigma) according to the manufacturer's instructions.

Teratoma formation. Undifferentiated human iPSCs were obtained by collagenase IV dissociation and injected subcutaneously into NOD/SCID mice. Palpable tumours were observed 4–6 weeks after injection. Tumour samples were collected at 6–8 weeks, cryo-preserved, processed by 10-μm cryo-sectioning, and stained with haematoxylin and eosin.

Karyotype analysis. Karyotype analysis was performed using standard G-banding chromosome analysis by the institution's cytogenetic laboratory according to standard procedures.

Bisulphite sequencing. Genomic DNA (1 μg) was bisulphite converted with the Methylamp DNA Modification kit (Epigenetek) according to the manufacturer's instructions. After bisulphite conversion, DNA was amplified using Faststart Taq polymerase (Roche) as follows: 94 °C for 4 min; 5 cycles of 94 °C for 30 s, 59 °C for 3 min, 72 °C for 3 min; 35 cycles of 94 °C for 30 s, 62 °C for 30 s, 72 °C for 30 s; 72 °C for 10 min. Primer sequences can be found in Supplementary Table 2. PCR products were TA-cloned into pTZ57R/T plasmid (Fermentas). Inserts were sequenced with M13 universal primers.

Gene expression analysis. Undifferentiated human iPSCs and differentiating embryoid bodies were frozen in liquid nitrogen. RNA was isolated using the RNeasy-plus mini-kit (Qiagen-Gm). Reverse transcription into cDNA was conducted using the high-capacity cDNA reverse transcription kit (Applied Biosystems). The PCR-related primers are detailed in Supplementary Table 2. Briefly, each RT–PCR included the following PCR program: 3 min at 93 °C, 30 s at 93 °C, 30 s at 60 °C, and 30 s at 72 °C. 2.5 ng of cDNA was used from each sample.

SYBR-Green real-time PCR studies were performed using the Fast SYBR Green Master mix (Applied Biosystems) and primers (detailed in Supplementary Table 3). All real-time PCR experiments were conducted in triplicates. Samples were cycled 40 times using a Fast ABI-7500 Sequence Detector (Applied Biosystems). ABI-7500 cycle conditions were as follows: 2 min at 50 °C, 15 min at 95 °C followed by 40 cycles of 15 s at 95 °C, 30 s at 60 °C and 30 s at 72 °C. Cycle threshold (CT) was calculated under default settings for real-time sequence detection software (Applied Biosystems).

Whole-cell patch-clamp recordings. Spontaneously contracting embyoid bodies were mechanically isolated, enzymatically dispersed into single cells (1 mg ml^{-1} collagenase B; Roche) and attached to fibronectin-coated glass coverslips. Action potentials were recorded from spontaneously contracting small clusters superfused with Tyrode solution at 32 °C under spontaneous and paced (0.5–1 Hz) rhythms. The Tyrode solution consisted of NaCl (140 mM), KCl (5.4 mM), $CaCl_2$ (1.8 mM), $MgCl_2$ (1 mM), HEPES (10 mM) and glucose (10 mM), pH 7.4. The pipette solution consisted of KCl (120 mM), $MgCl_2$ (1 mM), Mg-ATP (3 mM), HEPES (10 mM) and EGTA (10 mM), pH, 7.2. Action potentials were recorded using the current-clamp mode using Axopatch 200B, Digidata 1322A, and pClamp 9 (Axon Instruments) for data amplification, acquisition and analysis, respectively. A custom-written computer program was used to measure APD_{50}, APD_{70} and APD_{90} (the time intervals required to reach 50%, 70% and 90% of repolarization), dV/dt_{max}, repolarization velocity and action-potential amplitude.

For voltage-clamp studies of I_{Kr}, external Na^+ was replaced by equimolar choline (126 mM) and the solution was supplemented by 4-AP (5 mM), $BaCl_2$ (0.5 mM), $CdCl_2$ (0.2 mM) and chromanol (10 μM) to suppress potential interference of I_{Na}, I_{to}, I_{K1}, I_{Ca} and I_{Ks}, respectively. I_{Kr} was defined as the E-4031-sensitive (1 μM) current and was elicited by 3-s depolarizing steps from a holding potential of −60 mV to potentials ranging from −40 to +60 mV in 10 mV increments. This was followed by a 2-s repolarization phase to −30 mV to elicit tail current.

Microelectrode array recordings. A high-resolution microelectrode array recording system (Multichannel Systems)[28] was used to characterize the electrophysiological properties of human iPSC-CMs. The contracting areas within the embryoid bodies were microdissected and plated on fibronectin-coated microelectrode array plates. The recorded extracellular electrograms were used to determine local FPD, defined as the time interval between the initial deflection of the field potential to its return to baseline. FPD measurements were either compared in rate-matched specimens or were normalized (cFPD) to the activation rate using Bazett's, Hodges' or Fridericia's[16] correction formulae. For Bazett's correction, $cFPD\text{-}B = FPD/(RR)^{1/2}$; for Hodges' correction, $cFPD\text{-}H = FPD + 0.105 \times ((1/RR) - 1)$; for Fridericia's correction, $cFPD\text{-}Fr = FPD/(RR)^{1/3}$, where RR indicates the time interval (in seconds) between two consecutive beats.

To assess for the effects of different drugs on the electrophysiological properties of the cells being studied, between 5–30 μl of tested drug stock solution was added to 1 ml of medium to achieve the necessary concentration. The tested drugs included isoproterenol hydrochloride, nifedipine, ranolazine dihydrochloride, pinacidil monohydrate (all from Sigma), cisapride (Janssen-Cilag) and E-4031 (Alomone Labs). Preliminary studies (data not shown) demonstrated that DMSO (solvent for nifedipine and pinacidil) alone did not have any effect on the measured electrophysiological parameters. Extracellular recordings were performed for 60 s at baseline and after 5 min following drug application.

Statistical analysis. Results are reported as mean ± s.e.m. Comparison between LQTS human iPSC-CMs and control healthy human iPSC-CMs groups was performed using the unpaired Student's *t*-test. Drug effects on the cells were analysed by paired *t*-test. One-way ANOVA followed by Tukey's post-hoc tests were used when comparing multiple groups. $P < 0.05$ was considered statistically significant.

25. Lowry, W. E. *et al.* Generation of human induced pluripotent stem cells from dermal fibroblasts. *Proc. Natl Acad. Sci. USA* **105,** 2883–2888 (2008).
26. Arbel, G. *et al.* Methods for human embryonic stem cells derived cardiomyocytes cultivation, genetic manipulation, and transplantation. *Methods Mol. Biol.* **660,** 85–95 (2010).
27. Park, I. H. *et al.* Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451,** 141–146 (2008).
28. Kehat, I., Gepstein, A., Spira, A., Itskovitz-Eldor, J. & Gepstein, L. High-resolution electrophysiological assessment of human embryonic stem cell-derived cardiomyocytes: a novel *in vitro* model for the study of conduction. *Circ. Res.* **91,** 659–661 (2002).